**前前言：**根据2018版的稍加一些简易的修改，主体仍为2018版。注意画横线的句子，为曾经的考题。建议配合PPT一起食用。包含了2023年与之前学长姐的期末考题（划线句）

**前言**：2016级整理的复习资料。这个Word文档基本就是PPT的粘贴版本了，内容非常详细，但排版很乱，所以索性上传Word版本，学弟学妹们可以直接改动。本人生信最终得分91，失分的地方应该是填空和简答。选择和判断应该没错，关于选择中Python的6分，个人推荐MOOC上山东大学的Python教程，2倍速，没时间的话只看前4章就行，有时间全看完。个人觉得关于简答的最后一题非常灵活，这就考验你上课有没有认真听讲了。总的来说，生物信息学这门课程非常值得学习，老师上课水平很高，也很有意思，希望学弟学妹们认真对待，别留遗憾。

CHAPTER1

生物信息学就是利用信息技术对生物信息进行获取、储存、分发（查询）和分析，以解释这些信息数据所蕴涵的生物学意义的学科。

－－生物信息的产生源于自然选择

信息的“行为方式”——传递：一种选择引起另一种选择。包括复制传递和转化传递。

“为什么说“大分子的序列决定结构，继而决定功能”这一自然原理对生命现象之存在和演化有着重要意义”

GPT4答：生物信息的产生源于自然选择：在生物进化过程中，自然选择作用于个体的表型，而表型的差异往往由分子层面的差异所决定。个体的DNA序列决定了其蛋白质的序列，进而决定了蛋白质的结构和功能。那些具有有利表型的个体，其分子层面的特征更有可能被保留下来并传递给后代。因此，序列、结构和功能的相互关系是自然选择作用的基础，是生物多样性和复杂性的源泉。信息的传递方式：在生物体中，信息的传递主要通过两种方式：复制传递和转化传递。复制传递指的是遗传信息（DNA）的复制和遗传给后代，保证了生物信息的延续。转化传递涉及到信息从DNA转录到RNA，再从RNA翻译到蛋白质的过程。在这一过程中，序列信息被转换成结构信息，最终转换成功能。这一转化过程不仅是生物体内部信息流动的基础，也是生物体对环境变化作出响应的基础。选择引起选择：在生物进化中，一种选择（如环境压力）常常引起另一种选择（如生物体的适应性变化）。这个过程在分子层面上反映为序列的变化（如点突变、重组等），这些变化可能导致结构和功能的改变，从而影响个体的适应性和生存率。例如，某种酶的活性中心的微小改变可能显著影响其催化效率，进而影响个体的代谢率和生存能力。这样的分子层面的变化，经过自然选择的筛选，可能成为种群适应环境的重要因素。

生物信息学研究内容：

从信息技术来看

上游：有针对性的计算机技术的开发

中游：利用这些技术建立数据库、算法、软件和工具

下游：利用这些工具有效管理和处理生物学数据

上游：1.数据库管理技术。

2.数据仓库、数据挖掘与知识发现技术。

3.分布式计算（包括网络计算）。

4.图像处理和可视化技术。

中游：1.数据库的构建。

2.算法建立。

3.统计模型建立。

4.工具软件开发。

下游：1.数据库检索与数据挖掘。

2.建立特定方向或自己的专用数据库。

3.数据分析：序列分析、结构分析、进化分析等

从生物信息来看

本身、产生、传递三个方面

目前主要包括：

macromolecular sequences; （最初）

macromolecular structures;

expression profiles; (EST; microarrays; 2D-PAGE)

biochemical network; (Interactions and reactions)

evolution history.

HGP：1990～2000年，10年时间实现了“工作草图”，2003年实现了“完成图”，3×109个碱基对（30亿美元），并对约30,000个基因进行了注释。越来越多的其他模式生物也完成了全基因组测序工作。

另一个和GenBank基本等价欧洲核苷酸序列数据库EMBL的发展情况（1990-2003），也是明显的这种指数上升。人类基因组计划后，现在仍是这种指数增长的趋势。

生物信息学的产生是因为生物信息的全面发掘——巨大数据量的产生。计算机与信息科学的迅猛发展——为巨大数据量的处理提供了可能。

前基因组时代：生物数据库的建立、检索工具的开发以及DNA和蛋白质序列分析。如Genbank，Sequence alignment算法等。 (1980s)

基因组时代：基因组测序，基因寻找和识别、网络数据库的建立和交互界面的开发等。如基因组信息组装，建立与发展EST（Expressed sequence tag，表达序列标签）数据库等。(1990s)

后基因组时代：大规模基因组分析、蛋白质组分析及各种数据的比较和整合。例如，蛋白质组学（Proteomics）的产生以及功能基因组学的发展等。 （系统生物学 Systems biology）。(2000s~…)

The past decade has witnessed a staggering increase in the abundance and availability of molecular data produced, for example, from experiments in gene expression, proteomics, and DNA sequencing.

在过去的十年中，分子数据的丰度和可用性出现了惊人的增长，例如，基因表达，蛋白质组学和DNA测序中的实验。

大数据时代：异质数据（不在同一管理的数据库中），人工查询效率极低，需要利用计算机工具自动收集重要信息---数据挖掘。

生物信息学面临挑战：

建立生物信息系统相关的数据挖掘技术；

发展揭示大规模数据集合中不同组分之间关系的统计分析方法及优化算法.

开发各种类型的数据转换工具；

功能基因组学研究（系统生物学的深入探索）；

蛋白质相互作用网络；

基因组非编码区信息挖掘.

蛋白质结构的预测；

唐力、阿卡波糖——糖尿病 Nucleic Acid Research (一月数据库专辑； 七月软件资源专辑)

CHAPTER2

数据管理方式的发展：手工 文件系统 数据库

数据库（Database, DB ）：统一管理的相关数据的集合。（相当于一种文件格式）

数据模型（Data model，DM）：数据库结构和语义的一种抽象描述。

数据库管理系统（DB management system，DBMS）：对DB进行管理的软件，可提供DB的建立、查询、更新以及各种数据控制功能。（相当于管理文件的软件。例如docx为一种带格式的文本文件，而Microsoft Word为管理docx的软件）

3个抽象级别：内部级、概念级和外部级

内部级也称物理级，指数据以什么样的数据结构存储在物理硬件上。概念级也称逻辑级，指数据库管理系统本身依照什么逻辑管理数据，不再考虑物理硬件。外部级，指DBMS呈现给用户的方式，一般是表格式呈现。

两级数据独立性：物理独立性－－内模式的改变，通过修改映射2，可使概念模式以上的内容不变。 逻辑独立性：如果概念模式变了，通过修改映射1，可使外模式不变。

数据模型（Data model，DM）：数据库结构和语义的一种抽象描述（由数据结构、数据操作和完整性约束三部分组成）(通常用来标征数据库的类型)。

根据数据模型分类，

根据其所采用的数据模型的特性，数据库分为：

第一代：层次、网状 （1970s－1980s）

第二代：关系 （1980s－ ）

第三代：面向对象、对象-关系

关系型数据库目前应用仍最为广泛 —— 全球各大数据库系统厂商的主流产品都是针对关系型数据库的，如 DB2、Oracle、Sybase、MySQL、SQL Server 等。（SQL发音为/ˈsiːkwəl/中文谐音C寇）

上述各产品都可采用SQL (Structured Query Language)

—— 此语言简练而自然、易学易用，被ISO定为关系数据库语言的国际标准。

<表示5‘端；>表示3’端。

DBMS：以前是一个程序处理一组数据，现在是所有的程序都可以利用这一组数据，但程序和数据间的联系要通过DBMS这种专用程序来进行（DBMS还负责用户和数据库的直接交流）——即，DBMS是所有应用利用数据库的统一界面（或称为媒介，intermediary [,ɪntə'miːdɪərɪ]）。

数据库类型的概念上由数据模型来定义，实际上则是由DBMS来决定。

文件系统出现，转向数据处理：计算机应用从过去单一的科学计算转向了以数据处理为主，从而使计算机进入了各行各业乃至普通家庭。 因为在现实中，计算机的最主要应用在于数据处理（而不是计算，日常计算只需要计算器就够了），如文字处理、表格处理、图像处理等。其实，直到现在，我们的pc机仍然是基于这种数据管理方式。 然而，有一些情况进一步注重数据，不仅是一个程序处理就行了，需要专门的数据管理，供多方面的应用——即以数据为中心——数据库的出现，如标本馆，图书馆，银行系统，联网售票系统等等（大家可以考虑还有什么别的例子）。当然，我们要讲的生物信息数据系统也是这样。

数据库——数据内容上和形式上都有直接的关系----数据管理的代表

数据仓库——数据通常只有内容上的关系而且不一定是直接的---数据挖掘技术的一种前期步骤

数据库（Database, DB ）：统一管理的相关数据的集合。

数据仓库（Data warehouse） ：面向主题的、集成的、持久的、历史的数据集合。在集成到数据仓库之前，原始数据可能驻留在许多不同的数据源中，以文本文件、层次型数据库、关系型数据库等或它们的混合面目出现。

计算机网络是计算机技术与通信技术相结合的产物\*，包括计算机软/硬件、网络系统结构以及通信协议等内容。（网络系统结构，即网络分层模型，例如TCP/IP模型，OSI分层参考模型。通信协议，更准确的说是网络协议(protocol)，例如常见的HTTP，即HTTProtocol）

Internet是当今最大的物理网络系统，其中的World Wide Web (WWW)是Internet上的一个庞大的互联信息网络（更准确的说Web是Internet上的一个应用，这个应用负责储存并传输网页，其余的应用例如SMTP邮件应用，FTP文件传输应用）。Internet——现今世界上最大的广域网。WWW实质上是Internet提供的基于超文本信息组织结构的一种信息服务。（超文本信息，即HTT）

Internet——物理网络；W W W——信息网络 （应用）

TCP/IP（Transfer Control Protocol / Internet Protocol）是Internet 使用的（两个）基本通信协议。

TCP负责打包和还原（并按IP协议格式交付IP协议，以及从IP模块将IP传输的信息还原为适合交付给应用的信息），IP负责传输到位（将网络包从一个主机传输到另一个主机） 。

具体来说，TCP/IP协议包含分布在多个层次（从物理层到应用层）上的上百个各种功能的协议，我们熟悉的如超文本传输（信息浏览）HTTP、文件传输FTP、电子邮件SMTP和远程登陆Telnet等。

IP地址实现了单一计算机在整个internet中的定位。 路由器根据目的地的IP地址对消息进行转发。

URL(Uniform resource locator)统一资源定位器（符）URL实现了单一文档在整个WWW中的定位 例如 http://www.lib.whu.edu.cn/dzzy/dzzy.htm

Internet大大改变了生物信息管理和利用的面貌。

Internet的“适时产生”为生物信息学的迅速兴起提供了契机。

WWW是生物信息学资源共享的主要界面。

算法是为了解决一个特定问题而需要执行的一系列步骤或指令。正确性和效率

模型是为了描述或研究一个特定系统，通过抽象和简化而建立的反映这个系统中各种关系的一种形式。

总的来说，算法是一种解决问题的策略，而模型是一种所研究对象的描述。

然而，我们可以看出它们实际上是有一定相似性的，从表观上来看～～都需要输入数据，且都能够输出数据。

算法是根据输入数据得出输出数据——问题的解。

模型是根据输入数据得出输出数据，从而反映其所代表系统的特性。

正是由于这种相似性，算法和模型的概念可以相通，

一些成熟的模型（不再是为了描述或研究原系统）根据输入数据得到的输出数据如果可以作为问题的解，则这种模型就可以作为解决问题的一种策略，也可以说是一种真正意义上的、可以直接应用的算法，如隐马模型、人工神经网络模型(狭义人工智能)等等（遗传算法也具有类似特性---即利用遗传和进化的模型形成算法的思想）。

广义的人工智能指机器学习，人工神经网络，深度学习为狭义人工智能

MS-Windows（DOS）: Home and office PCs;

Unix: Workstation and servers;生物信息学的主流操作系统 （更主流的是Linux和macOS，后者为类Unix系统）

稳定性好: Over 30 years in industry and academia.

开放性好：Supporting possible tasks in future.

Internet上的操作系统：The software that powers the Web was invented in Unix, and many if not most web servers runs on Unix servers.

科学软件的载体：Many good-quality, interesting and important scientific softwares are written for Unix.

共享的乐园：Many programs can be downloaded and installed on Unix systems for free.

——几乎所有的大型生物信息数据库都运行于Unix之上，如Genbank和EMBL。

Linux is a free, open source version of Unix.

MacOS: Apple Macintosh

Windows x-window MacOS X

Dos Unix(-like) Unix(-like)

计算机语言分类

程序（programming）语言：C/C++, Pascal, Java\*, ……（原代码汇编或编译为可执行程序而执行）（注意Java不是直接编译成可执行程序，而是编译为字节码，由Java虚拟机进行二次解释）

脚本（scripting）语言：MS-VB, JavaScript\*, PERL, Python, R, ……（原代码作为“母”程序的数据被调用执行）

置标（markup）语言：HTML\*, XML\*, ……（超文本描述）

Perl主页骆驼

R and MATLAB are high-level languages with extensive built-in mathematical and statistical capabilities and many libraries.

Perl 和Java直接面向应用。

Perl 是 loosely typed 语言，语法和参数多种多样，给了程序员很大的自由，大大缩短了程序开发时间。然而，不适合模块化开发 ……Perl是一门自由且功能强大的编程语言 Bioperl 是一组 Perl 模块，它主要目的在于利用 Perl 解决生物学研究中的一些问题，主要是生物信息学中的各种实际问题，如获取分子生物学数据，分析序列文件，序列间比对，大批量 BLAST ，数据挖掘，等等。

Java 是 strongly typed 语言，在程序设计上就把程序员禁锢在一个保险箱里,使他们避免犯错误。适合于模块化开发（软件工程）……

C & C++ 常被用来开发底层程序，编程所需考虑的更多（如内存管理），但执行效率更高 灵活度上，C++类似于java。

Python 的简洁之处，没有奇异的符号；无‘；’无｛｝

用缩进控制程序块，似乎有些掣肘，但强迫用这种格式有好处---程序清晰，一目了然。

Perl语言简介

@：数组----类似于c中数组的概念，但单元又没具体长度限制。

$： 变量

%：hash---不同于数组，用字符串索引，而不是以数字索引

例子

Python版本---实际上第4行应变为下面两行，但为了简单讲解，保持不变。

string1="".join(phrase)

print("\n", string1)

CHAPTER3

信息资源的指数增长：

——以数据库形式管理。 ——以网络访问形式共享

二次库是基于源库数据提炼分析及归纳的结果，如非冗余化，序列模式，结构分类，进化关系等。

注意，下面三个是核酸序列数据源库，不是二次库。

最重要的公共核苷酸序列数据库——

GenBank：<https://www.ncbi.nlm.nih.gov/genbank/> （美国） 所属机构 NCBI NCBI是最著名是数据中心，Genbank是其中最著名的数据库

EMBL：<https://www.ebi.ac.uk/> （欧洲） EBI

DDBJ：<https://www.ddbj.nig.ac.jp/index-e.html> （日本） NIG

数据交换，每日同时更新序列资料，所以无论是投送数据还是查询数据都基本等价。以GenBank为例——

GenBank flatfile (GBFF) 是GenBank数据库的基本信息单位。是三个等价数据库交换数据时所采用的格式。

DDBJ flatfile 与GBFF格式相同。

EMBL flatfile 的格式每行都带有两个字母的前缀，表明本行的信息类型，在内容上与GBFF相同。

RefSeq (Reference sequence)——核酸信息二次库。（可以类比为consensus序列。源库有很多不同的序列，而二次库中仅有几个典型的）

以下为GPT4的回答。

**RefSeq（参考序列数据库）和GenBank都是由美国国立生物技术信息中心（NCBI）维护的核酸序列数据库，它们提供大量的遗传信息，但存在一些关键区别：**

**来源和策划：**

**GenBank是一个综合数据库，收录了全世界科研人员提交的序列数据。它是开放的，任何人都可以提交序列到GenBank，这意味着它的数据量庞大且多样，但可能包含一些重复或质量不一的数据。**

**RefSeq提供的是经过NCBI策划和验证的参考序列。它旨在提供一个准确、经过验证的序列集合，用于作为研究的基准。RefSeq的序列通常是从GenBank的大量数据中筛选出来的，并且会经过专家的审查和注释。**

**注释和更新：**

**RefSeq数据库中的序列有着更加详细和经过校对的注释，包括基因结构、功能以及已知的变异等。这些注释会定期更新，以反映最新的科学研究。**

**GenBank的序列由提交者提供注释，NCBI不会对这些注释进行修改或更新，除非提交者本人进行更新。**

**序列的类型和组织：**

**RefSeq提供多种类型的序列，包括基因组、染色体、蛋白质等，并且为了方便用户，会将相关的序列（如一个基因的不同转录本）组织在一起。**

**GenBank则更像是一个巨大的序列存储库，包含了从单个基因片段到完整基因组的各种数据。**

**稳定性和可靠性：**

**RefSeq的每个记录都会有一个独特的、不变的识别号，即便是序列更新了，这个识别号也不会变。这使得RefSeq成为科学出版物和生物信息学分析中可靠的参考资源。**

**GenBank中的序列随着时间和数据的更新可能会变化，包括其访问号（accession number）。**

Gene —— 基因信息二次库（和RefSeq的区别：聚焦gene——无非基因DNA序列，对于基因的描述更全面，如正规的基因名称、染色体位置信息、基因功能等信息（其序列信息链接RefSeq）。更像三次库。）（而Refseq则是全基因组序列，即使对于不是基因的区域）

规范的功能描述：Gene Ontology （GO）基因本体论Gene Ontology (GO): for a controlled vocabulary to describe genes and gene products; 基因按功能分类。

——本身不是数据库，而是一种类似于“协议”的东西。但有其术语数据库GO Annotation （GOA）database，在EMBL-EBI。

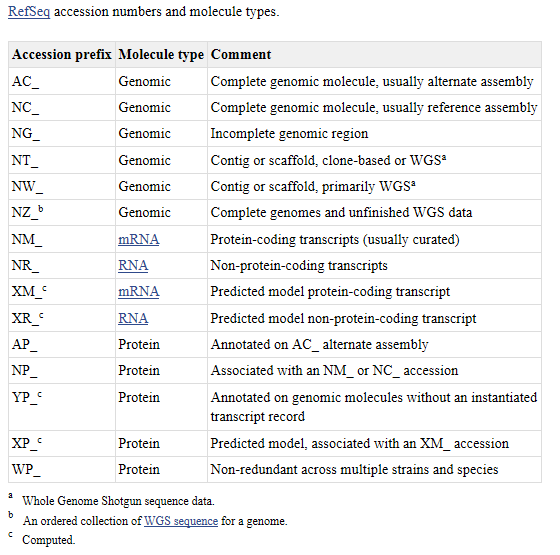
分三个大板块表述GO：

Molecular Function：由基因产物执行的分子级活动。分子功能术语描述在分子水平上发生的活动，例如“催化”或“运输”。GO分子功能术语代表活动而非执行动作的实体（分子或复合体），并且不指定动作发生的地点、时间或背景。分子功能通常对应于个体基因产物（即蛋白质或RNA）能够执行的活动，但有些活动是由多个基因产物组成的分子复合体执行的。广义功能术语的例子有催化活动和转运活动；狭义功能术语的例子有腺苷酸环化酶活动或Toll样受体结合。为避免基因产物名称与其分子功能之间的混淆，GO分子功能通常会附加“活动”这个词（例如，一种蛋白激酶的GO分子功能是蛋白激酶活动）。

Cellular Component：相对于细胞器和结构的位置，由大分子机器占据。基因本体论描述基因产物位置有两种方式：(1) 基因产物执行分子功能的细胞解剖实体。细胞解剖实体包括细胞结构，如质膜和细胞骨架，以及膜包裹的细胞隔室，如线粒体，(2) 他们是其部分的稳定的大分子复合物，例如，clathrin复合体。

Biological Process：由多种分子活动完成的更大过程或“生物程序”。广义生物过程术语的例子有DNA修复或信号转导。更具体术语的例子有嘧啶核碱生物合成过程或葡萄糖跨膜运输。注意，生物过程不等同于路径。目前，GO并不试图表示完全描述路径所需的动态性或依赖性。

NCBI Accession number： （N实验测得，X计算机预测）



还有NC开头的，是completed genome的意思，现在后面的数字扩展了，有的到达了9位（也有还是6位的）

最著名的基因组浏览器：NCBI Genome；UCSC；Ensembl —— 基因组浏览器。

蛋白质数据库——

序列： 源库 特征二次库

复合二次库 特征复合库

结构： 源库 分类二次库

蛋白质序列源数据库

PIR：1984年，美国国立生物医学研究基金会 （NBRF）建立，1988年收入MIPS（德国）和JIPID（日本），称为PIR-International或PIR-PSD（Protein Sequence Database）。 序列数据全面（多），但注释相对较少。已停用

Swiss-Prot：瑞士日内瓦大学建立。数据有详细注释，冗余度小，可靠性高，SWISS-PROT是最基本的蛋白质源数据库；TrEMBL：收集的是从EMBL中的核酸序列预测的蛋白质序列，体现了核酸测序优势对蛋白质研究的帮助。Swiss-prot：不像核酸数据，可能到refseq或gene库才有人专门审查和维护，蛋白质在源库水平就有专门的维护。这可能和核酸测序快，以及和功能距离稍微遥远有关。原来是PIR的二次库。现在，Swiss-Prot已经被包含在UniProt中。

NRL-3D：从三维结构数据库PDB中提取序列，其蛋白质三维结构均为已知。

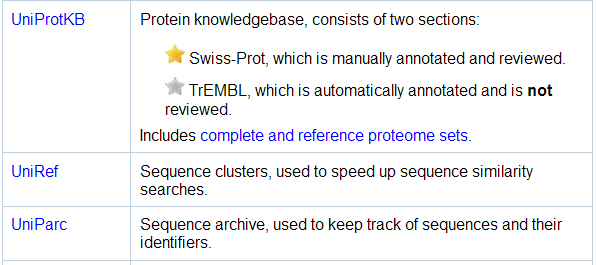
蛋白质序列复合数据库

——把不同数据库数据合并，并去除重复信息。

UniProt (Universal Protein Resource) is the world's most comprehensive catalog of information on proteins，joining Swiss-Prot, TrEMBL, and PIR. —— EBI，SIB (ExPASy) and PIR.

NRDB：由Swiss-Prot、 GenPept 、PIR 、NRL-3D 等数据库复合而成。—— NCBI.

UniProt分为三个层次：



蛋白质特征复合数据库—把不同数据库合并并去除冗余

InterPro

InterPro库中包含了大量关于蛋白质家族、结构域、重复序列以及功能性位点的信息。以蛋白质家族为例，如果你查询了一个特定的蛋白质序列，InterPro可能会返回以下类型的信息：

蛋白质家族（Protein Families）: 蛋白质家族是一组因功能、结构或序列相似性而归类在一起的蛋白质。例如，InterPro可能会指出某个蛋白质属于G蛋白偶联受体家族，这类蛋白质在信号传递中起重要作用。

结构域（Structural Domains）: 结构域是蛋白质中相对独立的部分，它可以独立折叠成稳定的三维结构，并通常具有特定的功能。例如，某蛋白质可能含有一个"SH2"结构域，这表明它可能涉及到磷酸化酪氨酸的蛋白质之间的相互作用。

重复序列（Repeats）: 一些蛋白质中含有序列重复，这些重复可能与蛋白质的结构稳定性或功能多样性有关。InterPro会指出这些重复序列的类型和位置。

功能位点（Functional Sites）: 功能位点是蛋白质上直接参与其生物学功能的特定氨基酸残基。例如，酶的活性位点或者配体结合区域等。InterPro可以帮助识别这些关键的生物学标记。

蛋白质结构源数据库

PDB (Protein Data Bank): 主要的蛋白质结构源数据库。信息来自实验如，X-ray diffraction, NMR，Cryo-EM. ——RCSB (Research Collaboratory for Structural Bioinformatics)

MSD (Macromolecular Structure Database) ——EBI （现在已与PDB没有区别）

2003年PDB+MSD+PDBj，形成wwPDB (Worldwide PDB)；2006年Wisconsin大学BMRB（Biological Magnetic Resonance Data Bank）也加入。

和蛋白质序列信息一样，蛋白质结构信息也愈来愈多。

蛋白质结构分类数据库

SCOP (Structure Classification Of Proteins)

依结构和进化的相关性进行分类，等级细化顺序为——Class，Folds、 Gene superfamily 、Gene family。SCOP (Structural Classification of Proteins) 数据库是一个广泛使用的蛋白质结构分类数据库。它的目标是通过考察蛋白质的三维结构来对它们进行分类，提供一个关于蛋白质结构域以及它们之间关系的详细框架。SCOP数据库的重点是识别和分类蛋白质结构中的模体（motifs），这些模体是蛋白质中具有特定功能和结构的独立单元。

CATH (classification by Class, Architecture, Topology, and Homology)

依结构和进化的相关性进行分类，等级细化顺序为——Class、Architecture、Topology、Homology superfamiliy、Sequence family。

还有其他的蛋白质数据库：双向凝胶电泳图： Swiss-2DPAGE，World -2DPAGE List……

蛋白质相互作用：BIND (Biomolecular Interaction Network Database), DIP (Database of Interacting Proteins), GRID (General Repository for Interaction Datasets)……

生化路径： KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway, UniPathway, EcoCyc ……

Interpro和SCOP的区别

InterPro:

集成性: InterPro是一个集成数据库，它整合了多个数据库的信息，包括蛋白质家族、结构域、重复序列以及功能性位点等信息。

功能和序列为基础: InterPro关注于通过识别蛋白质中的共有模式来预测蛋白质的功能和结构特征。这包括通过比对蛋白质序列到已知的家族和结构域等。

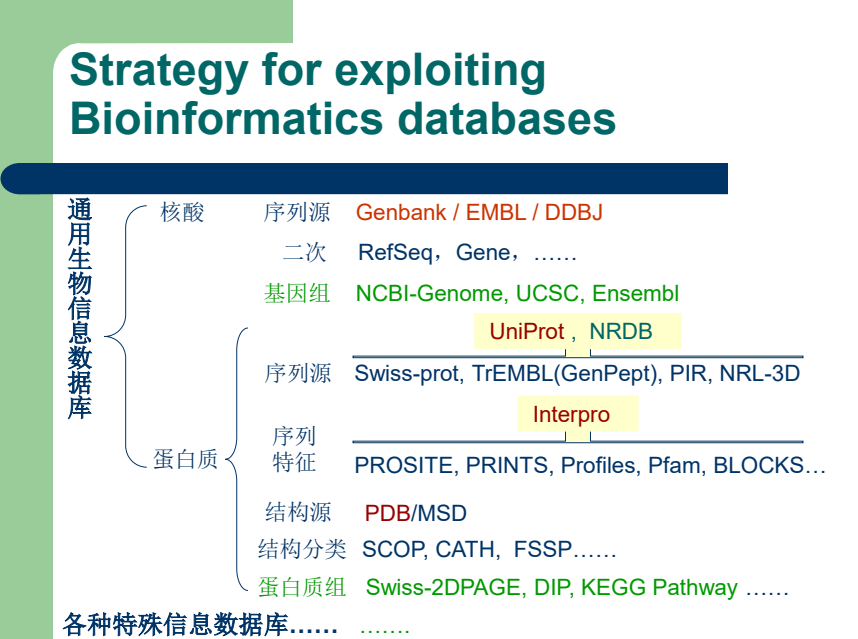
多样的数据源: 它集成了如Pfam, PRINTS, SMART等多个数据库的数据，提供了广泛的蛋白质特征信息。

SCOP:

结构为基础: SCOP的分类是基于蛋白质的三维结构，而不是序列。它将蛋白质按其空间结构的相似性进行分类。

手工分类: SCOP数据库中的蛋白质是通过专家手工进行分类的，这侧重于通过深入分析蛋白质结构的细节来进行分类。

层次结构: SCOP有一套详细的层次结构，包括类、折叠、超家族、家族等分类，这有助于研究人员理解蛋白质的结构和进化关系。



NCBI信息查询

关键词检索（Entrez\*) --- 注释信息

序列相似性搜索（Blast\*) --- 生物信息

**序列相似性分析的意义：**

**进化关系推测：序列相似一般是进化同源。原理：本质上，进化是趋异而连续的（逐渐变得不同）。相似性搜索即寻找同源序列。也有例外，如有些功能域的趋同进化，即相似却不同源，比如亮氨酸拉链中的coiled coil structure。 Instant note p79。还有趋异进化，即同源但不相似，如载氧的球蛋白家族，其中的成员有的之间只有10％的残基相同（但结构较保守）。－－－进化树的构建。**

**结构推测：如蛋白质结构预测中的同源建模（或模建）。**

**功能推测：在功能基因组学中有很大用处。**

**序列对位排列（sequence alignment）：源序列与目标序列之间按碱基(或氨基酸）位置相对排列；其目标是使序列之间的相似程度最大。**

相似性记分：相似性定量，如 A:A＝1；A:C＝0；A:-＝-1

…… 因此能得出记分矩阵。

**蛋白质序列相似性替代计分：考虑氨基酸特性：（也就是氨基酸之间的关系：如果是考虑进化方面，有密码子相似性；考虑结构和功能方面，有氨基酸性质）。**

蛋白质序列相似性替代计分：考虑氨基酸特性。

全局（global）排列：对序列全长进行最优对位排列。

局部（local）排列：通过对位排列使序列间的局部区域达到高度相似。

局部排列更具生物学意义，因为生物序列们往往只是局部高度相似，如两同源基因的外显子高度相似，而内含子区别很大。又如可能两蛋白质只有一个相近的域，而非全长相似。

序列相似性分析：

**已知序列的相似性比对：强调精确性。**动态规划算法，如Needleman-Wunsch算法（全局），Smith-Waterman算法（局部）。（动态规划算法的原理是将一个问题重复分解为可求解的子问题，在子问题上得到最优解后，原问题就可以得出最优解）

**数据库序列相似性搜索：强调速度**，原因是数据库大及WWW上的公众性访问。FASTA（启发式+部分动态规划，全局+局部，EBI）和 BLAST（启发式，局部，NCBI）。

BLAST局部比较，启发式算法，更快，对亲缘关系远的序列不敏感。with a minimal sacrifice of sensitivity to distant sequence relationships. 计分有明确统计学意义

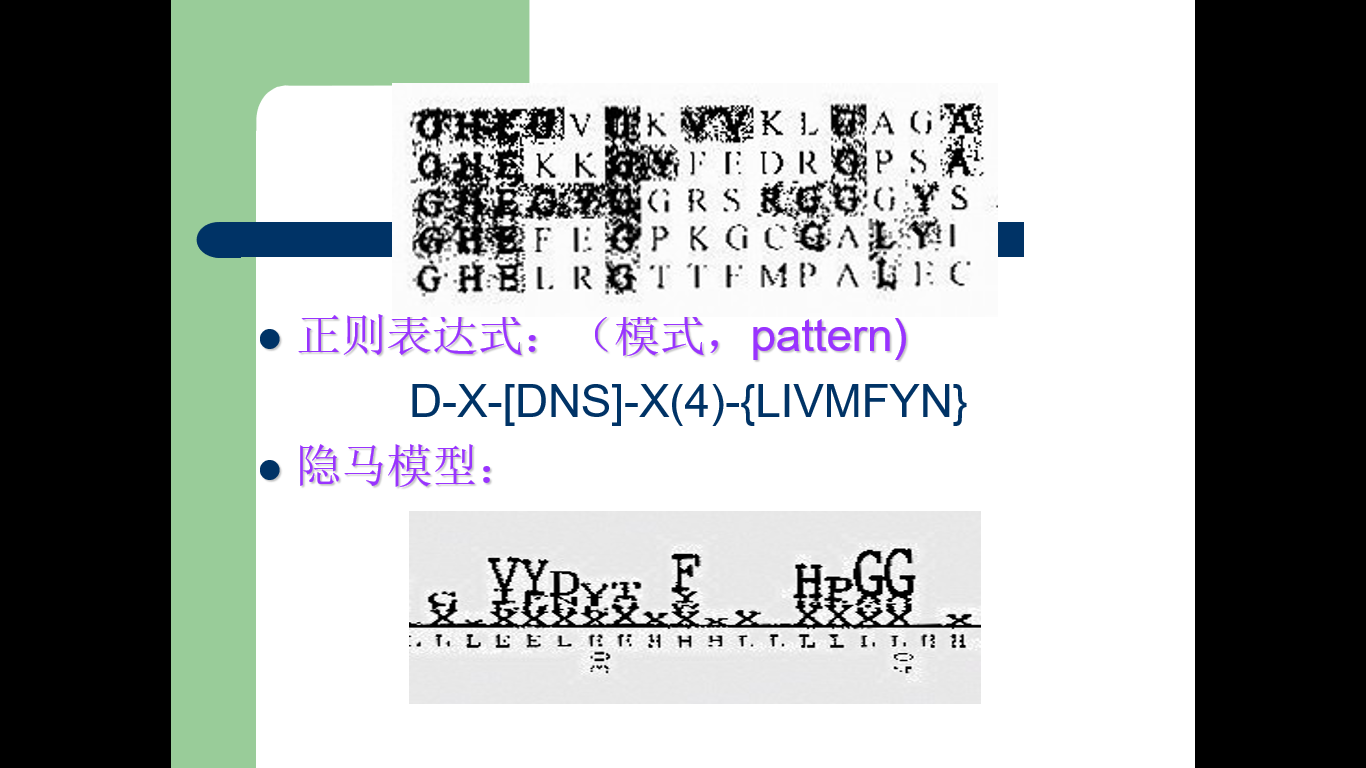
NAR的数据库专辑介绍实时更新的数据库的最新版本

信息获取的方式：浏览和查询。查询分为关键词查询和序列相似性查询

关键词查询：以名称、记录号、分类学等级、文献题目、作者等作为关键词提交，对数据库中的注释信息进行查询。

序列相似性查询：以序列本身作为提交信息对数据库中的序列信息进行查询。

PubMed: 最全的生物及医学文献题录库。（NCBI子库）



正则表达式：regular expression. 是用来描述文本特征（字符串）模式的，这里用来描述aa序列模式。

正则表达式，又称正规表示法、常规表示法(英语:Regular Expression，在代码中常简写为regex、regexp或RE)，计算机科学的一个概念。正则表达式使用单个字符串来描述、匹配一系列符合某个句法规则的字符串。在很多文本编辑器里，正则表达式通常被用来检索、替换那些符合某个模式的文本。

[]表示可以是其中任意一种，{}表示可以是除这些之外的任意一种。问题：共描述了几个位点？

隐马模型：一字母下的x，应表示除此aa外任意aa占比之平均值（此aa占比应加此平均值）；单独的x，应表示任意aa的平均值（之所以有高有低，主要因为x包括的种类有多少）

BLAST的主要思想：

启发式搜索：BLAST使用一种启发式方法来加速搜索过程。而不是查找数据库中所有可能的序列位置进行全面比对，它寻找短的单词或"种子"，这些种子在查询序列和数据库中的序列之间精确匹配或非常相似。这大大减少了比对的数量，因为只有在找到这些种子匹配的情况下，才会进一步扩展比对。

建立索引（不发生在检索时）：数据库在先前已经根据各个种子建立起了索引。即这个种子可以在数据库的哪条记录的哪个位置来找。用户输入序列后以种子匹配到后，再根据种子的索引进入特定的序列。

种子扩展：在确定种子后，BLAST尝试扩展这些匹配到更长的相似序列。这个过程涉及在种子两边添加更多的序列，并计算一个对齐得分，直到得分不再增加或增加得很慢。这样可以得到局部相似区域，称为高分段配对（High-scoring Segment Pair, HSP）。

评分系统：BLAST使用一个评分系统来评估序列之间的相似性，该系统考虑了匹配、不匹配、序列间隔和开放缺口的成本。它使用一系列的评分矩阵（如PAM或BLOSUM）来评分蛋白质序列的比对，这些矩阵基于生物学上已知的替换模式。

统计分析：对于每个找到的相似性，BLAST还提供统计分析，比如E值（期望值），表示随机情况下找到类似或更高分数的相似性的频率。E值越小，匹配越不可能是随机发生的，即越有生物学意义。

S 和 E的意义（重要）

The significance of a given alignment with score S is represented by the E (expect) value, the expected number of chance alignments with a score of S or better. (给定对齐得分S的显著性由E值（期望值）表示，它是指随机对齐中得分为S或更好的对齐数量的预期值。)

E=kmne^(-λS)

The E value decreases exponentially as the Score (S) increases.

The E value reflects the scoring system (k,λ) in use, the length of the query sequence (m) and the size of the database (n) .

S是一个指标，但并不充分，因为分数高除了可能代表比对情况较好之外，还可能是别的原因造成，如如原始分值很高，查询序列较长，库很大。就好像你们一次考试的成绩一样，630分分数是很高，但是可能是题目简单造成的。那么一个更可靠的指标是——

通俗地说：E value反映的就是一个信心指数，E越小，我们的信心越大。S越高（E越小），我们的信心越大，但这种信心可能被其它因素降低，如原始分值很高，查询序列较长，库很大——可能是“虚假”的高分。E值的计算给我们更可靠的信心。

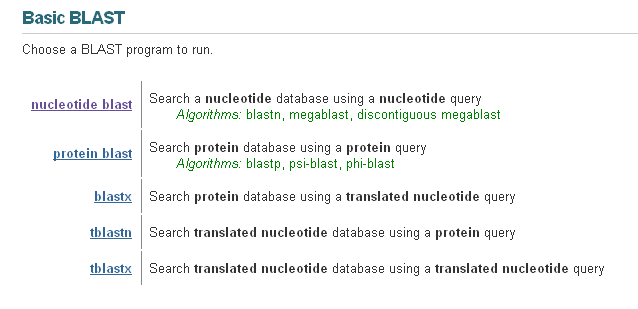
S值是一个绝对值，难以清楚评价命中的意义（如库太大，或打分矩阵的记分原始值太大），用统计学E值更清晰。

E越低，相似性越高。

特异性高错搜的少，敏感性高漏搜的少。

**敏感性也称真阳性率，特异性也称真阴性率。**

CHAPTER4



**多序列比对，直接比对很难。先两个比对。Clustal 用的就是progressive alignment。FastA格式输入序列，用phylip格式输出比对结果，为用phylip软件构建进化树作准备。**

**系统发育分析：**

**A phylogenetic analysis of a family of related nucleic acid or protein sequences is a determination of how the family might have been derived during evolution.**

**Two sequences that are very much alike will be located as neighboring outside branches and will be joined to a common branch beneath them.**

**The object of phylogenetic analysis is to discover all of the branching relationships in the tree and the branch lengths.**

对相关核酸或蛋白质序列家族的系统发育分析是确定在进化过程中如何衍生该家族。  
两个非常相似的序列将被定位为相邻的外部分支，并将连接到它们下面的公共分支。  
系统发育分析的目的是发现树中的所有分支关系和分支长度。

**由于遗传密码的简并性，蛋白质序列比DNA序列更加具有同一性。**

**一方面，进行DNA序列比对可能难以识别相似性较低的序列关系（不适合特别远缘的关系）；**

**另一方面，进行蛋白质序列比对可能丢失与进化过程直接有关的一些信息（不适合特别近缘的关系）。**

**蛋白质序列比对的局限性：**

**蛋白质序列比对通常更能反映远缘物种之间的相似性，因为即便是遥远的物种，它们的关键蛋白质功能区往往保守，即序列相似。但在比对特别近缘的物种时，使用蛋白质序列比对可能会丢失一些进化过程中的细微变化信息，因为即使DNA序列有变化，由于遗传密码的简并性，这些变化可能不会体现在蛋白质序列上。**

**DNA序列比对的局限性：**

**DNA序列比对更适合近缘物种之间的比对，因为它能够揭示微小的变化，这些变化可能在蛋白质序列中不明显。然而，当比对遥远的物种时，由于DNA序列变异较大，可能会难以识别它们之间的相似性。在这种情况下，由于进化时间的增长和随机突变的积累，即使是功能相同的基因，其DNA序列也可能显著不同。**

**比如，亲缘关系比较远的可优先使用？蛋白质比对，关系较近的优先使用？DNA序列。**

**比较亲缘关系远的应该用进化速率慢的，亲缘关系近的应该用进化速率快的。**

**概念性翻译：根据遗传密码表，理论上可以对任意一个DNA序列进行翻译而得到氨基酸序列，称为概念性翻译；这种通过计算机翻译而不是实验手段测定得到的蛋白质序列称为概念性序列或假设序列。**

**六框翻译：对任意给定的一段DNA序列，不知道其读码方向（即不知其是正义链还是反义链），也不能确定其编码区是否从第一个碱基开始，则必须将其所有的读码框全部都翻译出来，即——**

**先以所给DNA为模板，分别从（5’－3’）第1、2、3个碱基开始翻译，得到3种翻译结果；**

**再以其互补链为模板，依次从（5’－ 3’）第1、2、3个碱基开始翻译，得到另外3种翻译结果。**

**开放读码框（ORF）：起始于起始密码子、终止于终止密码子的“中间没有终止密码子”的读码框，称作开放读码框，又称可读框。**

**六框翻译一般只有一框是正确的，往往是其所包含的最长ORF最长的那框。**

长的ORF很可能意味着存在CDS，但往往需要进一步的审查；

对于特定物种，密码子有“偏爱”，可以用来鉴别这种预测的可靠性。

特例：一些基因组中GC丰富的物种，其密码子第三个碱基偏爱使用G或C，而不是A或T。

基因相关的非CDS特征序列预测

原核生物基因中，除CDS之外，还有很多重要的序列，如promoters 、ribosome-binding sites、termination sequences of transcription 等，它们大多是具有保守性的特殊序列，可以通过序列分析进行推测。

真核生物基因中又含有RNA剪接剪接事件，使得理论翻译与预测十分困难。另有借转录水平预测的（EST）

**综合性的基因搜索软件**

Glimmer （原核生物，准确度97％－98％）

GeneMark （原核生物、真核生物）

GenScan （真核生物——脊椎动物）

FgeneSH （真核生物，收费）

GRAIL （神经网络算法）

BGF（华大，水稻）

隐马模型简介

马尔可夫链（Markov Chain）是一种统计模型，用于描述一系列可能从一个状态转移到另一个状态的事件，其中下一个状态的选择仅依赖于当前状态，而与之前的状态无关。这个性质称为“无记忆性”或“马尔可夫性质”。简而言之，马尔可夫链可以用来预测一系列事件中下一步可能发生的事情。

马尔可夫链的组成：

状态空间： 一组可能的状态，系统在任一时刻都处于这些状态中的某一个。

转移概率： 从一个状态转移到另一个状态的概率。通常用转移概率矩阵表示

隐马尔可夫模型（Hidden Markov Model，HMM）是一种统计模型，它用于描述一个系统的马尔可夫链生成不可观察的状态序列，而这些不可观察的状态产生可观察的事件或符号。在HMM中，观察者无法直接看到状态本身，但能看到由这些状态生成的一系列观测数据。这个模型广泛应用于语音识别、生物信息学、时间序列分析等领域。

HMM的组成部分：

状态： 这些是模型的核心，代表系统可能处于的不同阶段或条件。这些状态对观察者是隐藏的。

观察： 每个状态可以产生一个观察，观察者可以看到这些观察，但不知道是哪个状态产生的。

转移概率： 每个状态转移到另一个状态的概率。

发射概率： 在某个状态下生成某个观察的概率。

初始状态概率： 系统开始时各状态的概率。

请注意，状态是没有马尔可夫性质的。

评估问题： 给定模型参数（给定转移概率矩阵）和某个观察序列，如何有效计算此序列的概率？这通常通过前向算法或后向算法解决。

隐藏状态和观测： 在这个模型中，DNA序列的每个部分（如外显子、内含子等）被认为是一个隐藏状态，因为它们不能直接从序列本身观察到。而DNA序列中的每个三核苷酸（即密码子）是观测结果。每个隐藏状态（比如外显子）都有一个发射概率，它定义了在该状态下观测到每个特定密码子的概率。

状态转移概率： 每个隐藏状态（例如，从一个外显子到内含子）还有一个转移概率，它定义了从当前状态转移到另一个状态的概率。

（以上内容在数学上仍是珍宝，但是实际应用中已不再使用，而转向使用深度学习神经网络）



蛋白质特征分析

"特征"（signatures）是一个广泛用来描述一系列有共同结构特点或序列相似性的蛋白质区域的术语。

InterPro是一个蛋白质序列复合特征二次库

Domains and motifs\* --- Interpro

一个结构域（模块、模块元素或折叠）是一种蛋白质区域，采用特定的三维结构。

一个基序（指纹、模式或轮廓）是一个短的保守蛋白质区域，通常为10~20个氨基酸 —— 不意味着同源性。

--- 一个结构域可能包含少量的基序，但一个基序不一定在结构域内部。

（domain和motif两个词语在事实上已经被混用了）

CDD（保守结构域蛋白序列数据二次库）

**CDD有自己提取的domain数据，属于特征二次库**

**ExPASy: PROSITE PROSITE: a dictionary of motifs**

**特性：如可被磷酸化等为特性。**

**高等真核生物往往编码序列大大 少于非编码序列，从基因组序列 出发的基因预测往往效率不高； 而从实际表达的基因着手可能有更高的效率。于是引出了EST。**

**EST（Expressed Sequence Tag）的意义**

**它们是从cDNA测序产生的短序列信息，可代表在特定组织或发育阶段表达的基因，其相关研究是一种发现新基因的有效方法。**

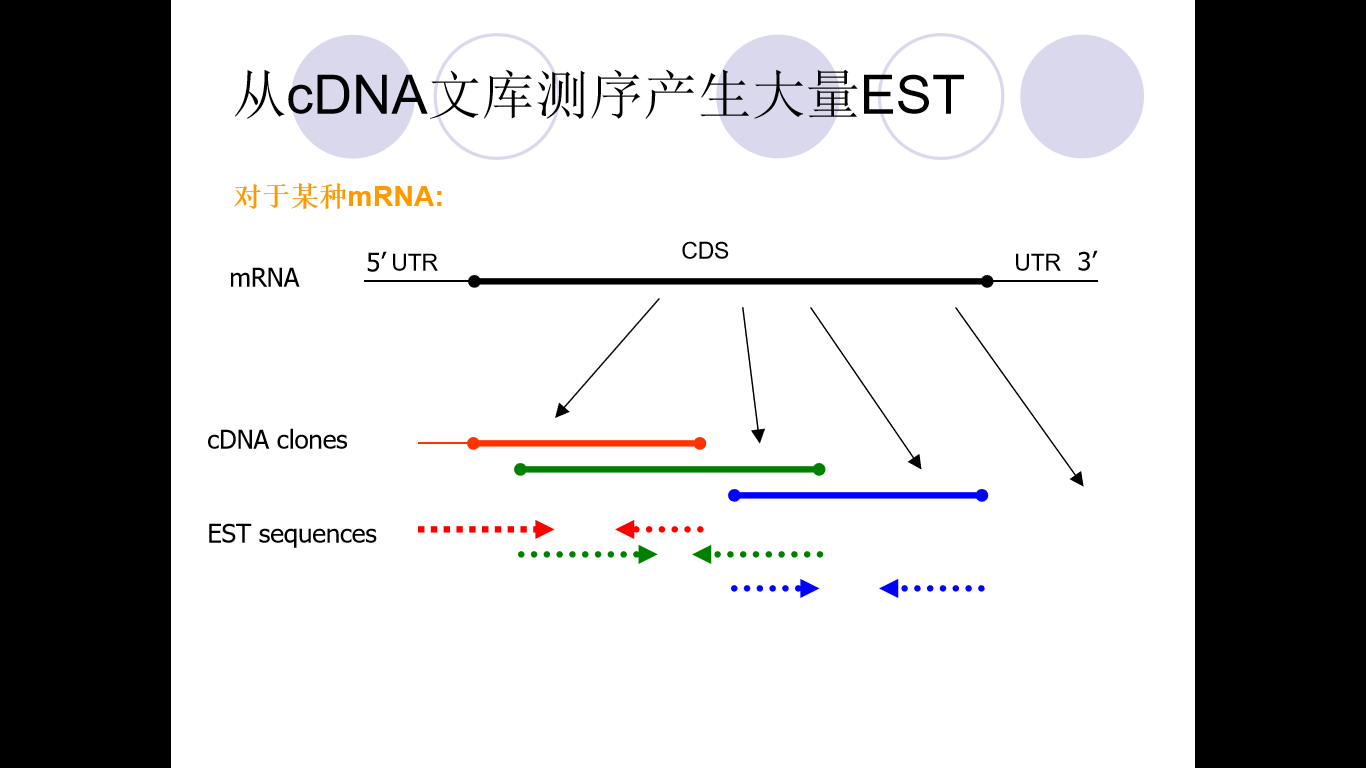
**cDNA文库：提取出某种细胞（或组织）的全部mRNA，在体外反转录成cDNA，重组入载体质粒（或噬菌体）后，转化受体菌（或包装） 。这样，每个受体菌（或噬菌体）都含有一段cDNA，可根据需要进行筛选或扩增，即对应一个cDNA克隆。这种包含一种细胞全部mRNA信息的cDNA克隆的集合称为此种细胞的cDNA文库。**

**-- cDNA文库中的某两个克隆，可能来源于同一种mRNA，也可能不是；可能是全长，但一般不是全长**

**cDNA克隆单次测序产生一个EST、对cDNA文库进行高通量自动测序：对每个cDNA克隆单次测序。由于其中cDNA片段的长度往往还是超过测序能力，这种测序往往只能产生一个片断序列信息，称为一个EST。**

**——也可以说，应用自动测序仪对一个cDNA克隆的一种“读法”产生一个EST。**

**这种读法可以从5‘端进行，也可以从3’端进行。**



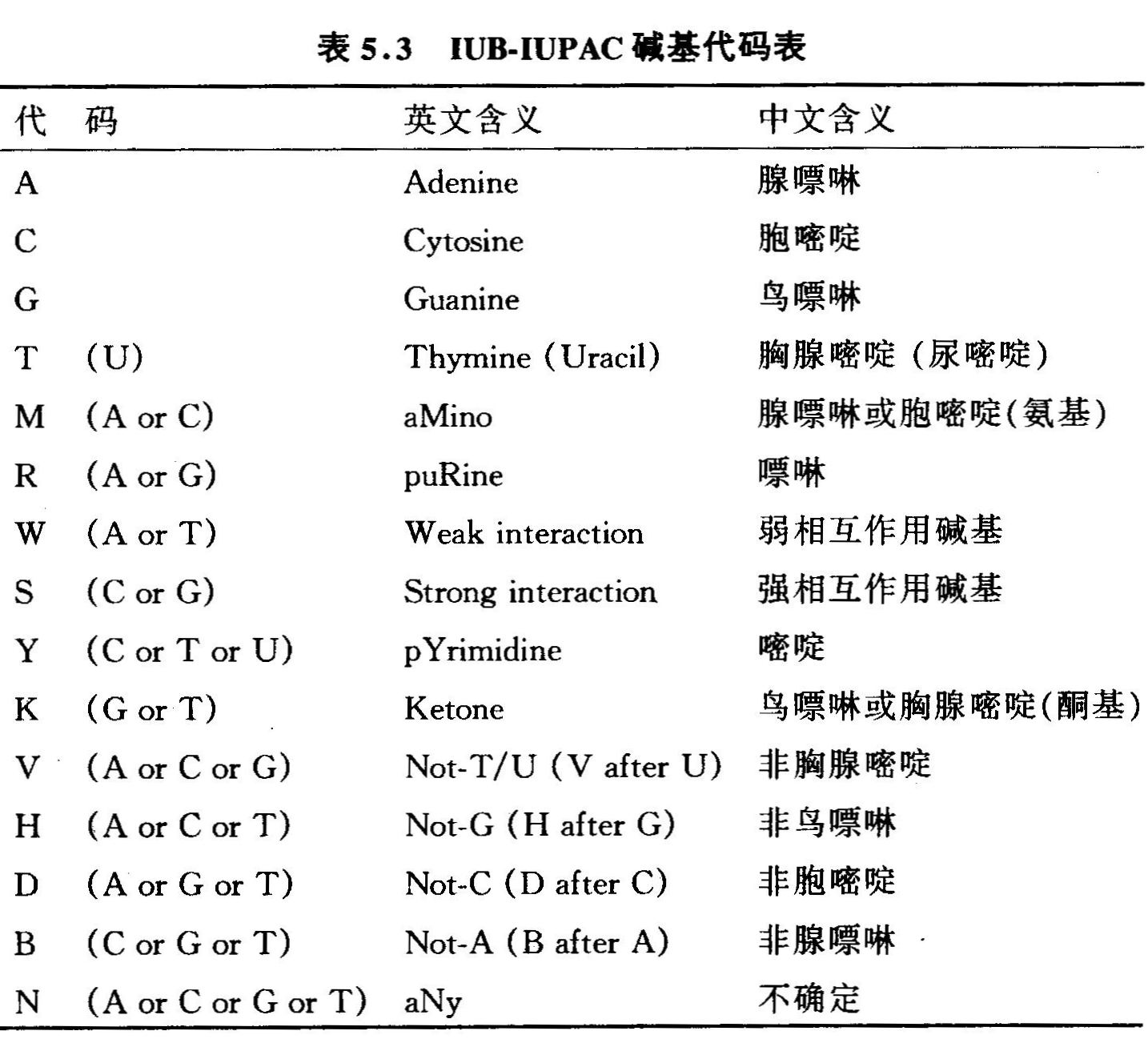
EST数据的特性

EST信息中除了A、G、T、C外，可能出现**模糊碱基**（如未知碱基N）EST为单次测序结果，经常出现模糊碱基；

EST信息可能出现**错误**,其中插入或缺失将导致翻译时读码框移位（frame-shifts）；

错误率：genome 1/10kb，EST 1/100.

低丰度表达的基因信息缺乏；而高丰度表达的基因在EST数据库中过量存在——高度冗余，即覆盖率过高，甚至一个EST序列可能是另一个EST序列的一个片断**。**



EST正在慢慢淘汰，RNA-seq技术和EST技术的最大区别只在“量”方面，前者量更大、

（事实上，EST已经被淘汰了，现在使用RNA-seq技术，EST的处理与分析手段与RNA-seq完全一致。）

EST相关分析

EST序列的预处理：去除原始EST序列中的载体、接头以及引物等“污染物”，得到干净的EST序列片段。

EST聚类：根据EST的重叠情况，判断同属一个基因的EST）EST自比对或将EST比对到基因组上。 Unigene is an ongoing effort at NCBI to cluster EST sequences with traditional gene sequences.

EST拼接：根据已经聚类的EST片段拼接，目标为全长cDNA序列。序列拼接软件——CAP3

phrap、Cap3、Cat、Staden、TIGR

**Domain较长，一般是反映同源性；**

**Motif较短，不一定反映同源性，可能是趋同进化的结果（由于功能的“需要”）。**

**（这段内容不考：对于RNA-seq来说常用的拼接与比对工具为HISAT2： 是TopHat的继任者，提供更快更有效的对齐。StringTie： 用于高效的重建和量化RNA-Seq转录本的工具，经常与HISAT2一起使用。）**

**CHAPTER5**

**什么是分子系统发生分析？**

**基本概念**

**同源性（\*直系同源）**

**类群 （\*单系类群）**

**系统发生树**

**基本原理**

**置换模型**

**序列分歧度**

**进化速率**

**分子系统树的构建方法（\*最大似然法）**

**分子系统发生分析软件**

基本概念：

**分子系统发生：A phylogenetic analysis of a family of related nucleic acid or protein sequences is a determination of how the family might have been derived during evolution.**

**Two sequences that are very much alike will be located as neighboring outside branches and will be joined to a common branch beneath them.**

**The object of phylogenetic analysis is to discover all of the branching relationships in the tree and the branch lengths.**

同源性（Homology） ：最基本的意义就是具有共同祖先，反映的是进化过程中的“亲缘关系” ( relatedness ) 。其说明的主要对象为“性状” (traits) —— 可以是宏观的形态和结构等，也可以是微观的大分子序列和结构等。

相似性（similarity）：…… 相似一般意味着同源，原理是——

“进化总体来说是趋异的，而且是连续的”。

（也有趋同进化……）

—— 用相似性来推断（determine）同源性

（当然，也可能是趋同进化导致的相似性，注意趋同进化没有同源性）

分子进化中的同源类型：

直系同源（orthology）：由“物种分化”而产生（常常有功能一致性），可以反映物种血统上的同源性，即物种进化的历史。

比如，人的肌红蛋白和小鼠的肌红蛋白。

并系同源（paralogy）：由基因“多重化（duplicating）＋功能分化” 而产生。

比如，人的肌红蛋白和其血红蛋白中的α、β亚基。

类群的定义

祖先类群（Ancestral group）:

祖先类群是指一个共同祖先及其部分后代的集合，但不包括所有后代。这个术语通常用于指代一个早期的物种或群体，它是后来更多分化物种的源头。（注意：祖先类群不一定是已经灭绝的。）

子裔类群（Descendant group）:

子裔类群是指源自某个共同祖先的所有后代。它通常用来描述由一个特定祖先进化而来的全部物种或群体。

单系类群（Monophyletic group）:

单系类群是指一个包含某个共同祖先及其所有后代的群体。这个概念是系统发育分类中非常重要的，因为单系类群反映了自然的进化关系。在单系类群中，任何成员间都比它们与外部群体更近亲。

并系类群（Paraphyletic group）:

并系类群包含了某个共同祖先和部分但不是全部的后代。这种类群遗漏了一部分真正的后代，因此并不完全反映了自然的进化关系。例如，爬行动物作为一个群体，如果不包括鸟类，就是一个并系类群，因为鸟类实际上是从爬行动物演化而来的。

复系类群（Polyphyletic group）:

复系类群是指一个群体，其成员来自于多个不同的祖先，而这些祖先本身并不包括在内。这样的群体并不反映共同祖先和后代之间的真实演化关系，而是由于表面相似性或者错误的解读而聚在一起。复系类群在现代系统发育学中通常被认为是不合理的，因为它们不基于共同祖先的概念。

将物种或类群之间的关系想象为一棵树，其中每个节点代表一个物种或类群，每条边代表它们之间的进化关系。

单系类群（Monophyletic group）:

在树上，单系类群可以被视为由一个节点（共同祖先）及其所有下游的节点（后代）组成的集合。在这棵树中，如果你选择了一个节点及其所有的子孙，那么你就选取了一个单系类群。这意味着如果你剪下树的任何一条分支，那么落下的部分（包括起点的那个节点和所有由它衍生出的节点）构成了一个单系类群。

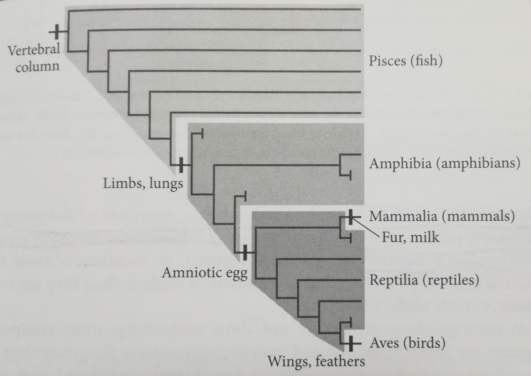
并系类群（Paraphyletic group）:

并系类群在树中表示为一个节点及其部分后代，但不是全部。这就好比你剪下了一部分树枝，但故意留下了一些枝条。在系统发育树上，如果你选择了一个祖先和一些但不是所有的后代，那么你就形成了一个并系类群。这种群体通常包括那些外表相似或传统上被分类在一起的物种，但缺少了一些重要的分支，这些分支可能已经发展出了显著不同的特征。

复系类群（Polyphyletic group）:

复系类群可以被视为树上非系统地选择的节点集合，这些节点并不共享一个直接的共同祖先，或者说它们的共同祖先并没有包括在内。在系统发育树上，如果你选择了几个彼此不直接相关的节点（通常是因为它们具有某些相似的特征或误解了它们的关系），那么你就形成了一个复系类群。这些节点可能分布在树的不同部分，不共享一个直接的共同祖先，或者其共同祖先很早就被排除在外了。

举例



陆生脊椎动物——单系类群。

鱼类：并系类群

温血动物（鸟和哺乳类）：复系类群，它不包含它们的共同祖先——如果不是强调这一点，则复系类群可能被认为属于一种并系类群。也就是说温血动物不包括羊膜卵类的共同祖先。

内类群（Ingroup）:

内类群是指在某个特定系统发育研究中主要关注的那些类群。在构建系统发育树时，研究者通常会选择一个或多个具有共同特征的物种或类群作为研究对象，这些被选中的对象就构成了内类群。内类群的成员被认为在某些特征上是相互关联的，研究者的目的通常是要解释这些内类群成员之间的相对关系和进化历史。

外类群（Outgroup）:

外类群是指与内类群相比较的参考群体。在系统发育分析中，外类群通常是已知与内类群有关联但又相对较远的物种或类群。选择合适的外类群是重要的，因为外类群可以帮助确定内类群的根，即最初的分支点。这样，研究者就能够判断内类群成员之间的进化关系，以及它们从共同祖先分化出来的相对顺序。

姐妹群（Sister groups）:

姐妹群是系统发育树上任意两个最近共同祖先相同的类群。在树状图上，姐妹群表现为一对在最近分支点分开的分支。这意味着它们是彼此最近的亲属，没有任何其他类群比它们之间的关系更近。在系统发育分析中，识别姐妹群有助于理解类群之间的具体分化事件和相对进化关系。

系统树(或进化树)：表达类群间系统发育关系（进化关系）的一种树状图

系统树一般是二叉树，是因为当分叉为2时表示一个特征分化成了两种。一般不存在一个特征同时分化成三种的情况，因此分化总有时间先后顺序（时间尺度可以很大，以年，十年，百年，万年记都可能），故而在每次分化只能是分化成两种。即二叉树。

构造n个节点的树，无根树的数量少于有根树。（只需考虑一个简单的事实，3个节点的无根树仅一种，而有根树中三个节点均能做根，故有3种。）

多歧分枝多为无法确定分化先后次序而形成，并非真正同时分化

标度树枝（scaled branch）系统树：树枝长度代表性状变异的数量。

非标度树枝（unscaled branch）系统树：树枝长度并不表示性状变异的数量。但其中节点的位置通常与分化时间相对应。

基因树（gene tree）：根据同源基因所构建的进化树。

物种树（species tree）：反映物种进化路径的进化树。

如何将多个（直系同源）基因树综合成一个物种树，是分子系统学目前所面临的一个难题。

（直系同源）基因树可能彼此不同，不一定是构树过程的问题，可能是进化过程本身造成……

核苷酸置换模型：进化的一个基本过程就是核苷酸随时间而变化（置换）。

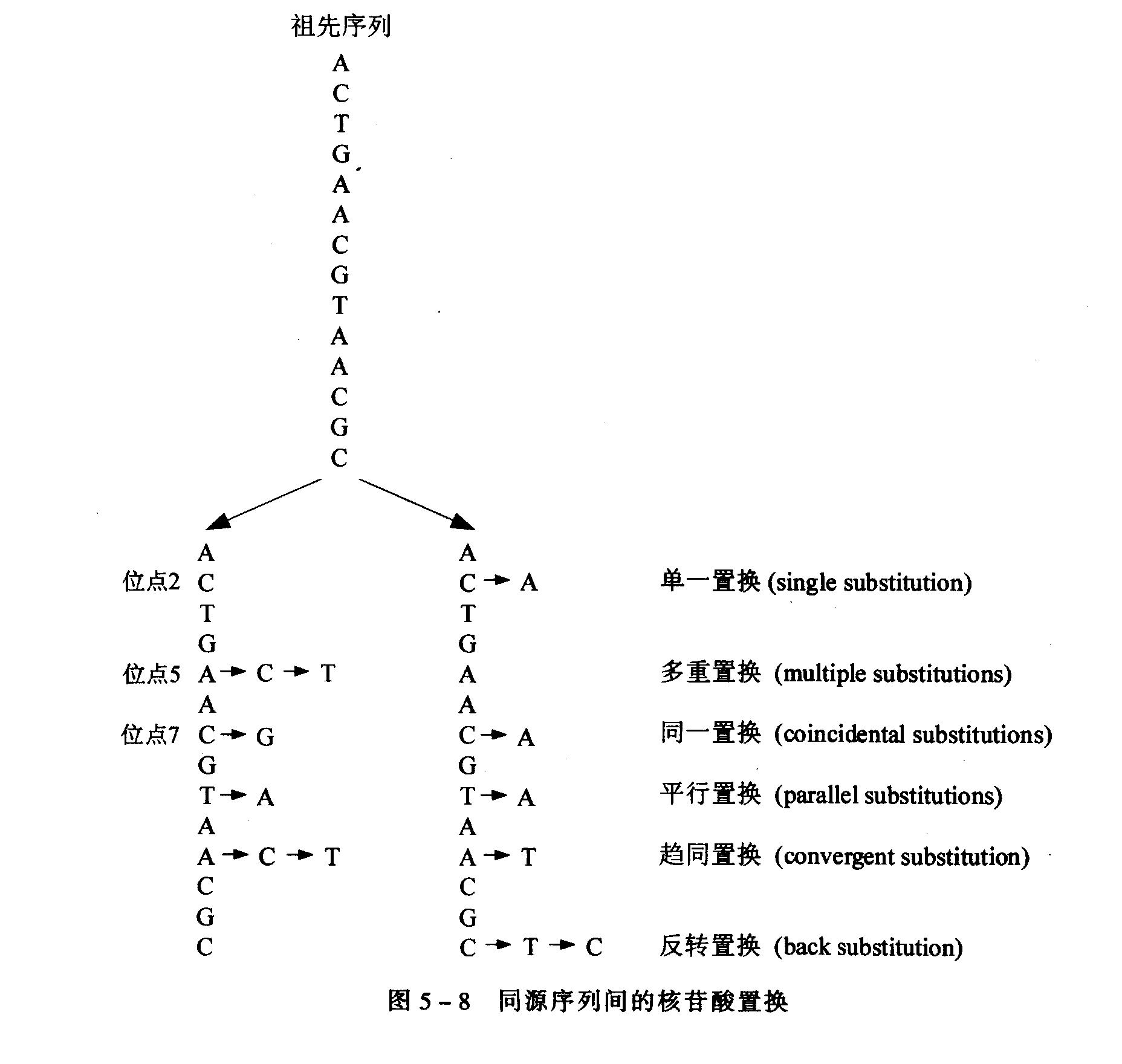
核苷酸置换模型可以用矩阵来表示从一个核苷酸到另一个核苷酸的置换概率。

氨基酸置换模型要考虑其密码子基础（即考虑变异），以及特性关系（如极性）（即考虑自然选择）。

根据这些概率可以推测物种之间的亲缘远近。例如两个物种之间核苷酸置换概率小的说明亲缘远，反之，则近。

置换种类

：



序列分歧度（sequence divergence）K是一种相异性指数。

设两个序列比对长度为L，残基差异值为N，则差异率P＝N/L，

DNA序列分歧度K是由差异率计算的一个值：

单参模型，K＝-¾ ln(1-4P/3) 不考虑转换和颠换区别,转换：DNA分子中的嘌呤被嘌呤或[嘧啶](http://www.so.com/s?q=%E5%98%A7%E5%95%B6&ie=utf-8&src=internal_wenda_recommend_textn" \t "https://wenda.so.com/q/_blank)被嘧啶替换。颠换：DNA分子中的嘌呤被嘧啶或嘧啶被嘌呤替换

蛋白质序列分歧度：

假设氨基酸之间的置换率相同，K＝-ln(1-P) 。

置换不一定带来差异

置换或相异性很大有时并不能表示亲源关系很远，可能是进化速率快造成。

进化分析要求序列相似性不能太小，或太大。

如构建人、黑猩猩和大猩猩的进化关系用到线粒体DNA中的一段非编码区。

构建真核生物、细菌和古细菌的进化关系用到核糖体RNA。

基于相似的原因，研究亲缘关系较远的进化关系用蛋白质可能比核酸好。

进化速率影响

变异本身的速率有差异。 RNA病毒：易错复制酶，如HIV、SARS病毒、流感病毒。 线粒体DNA：易错复制酶，高氧环境。（进化很快）

自然选择影响进化速率。 不同的分子（甚至域）具有不同的进化速率（保守性）。如非编码区进化很快，组蛋白和rRNA等重要基因进化很慢。

**距离矩阵法distance matrix method)**

首先，考虑多个序列组成的集合，简单地计算每两个序列的差异数量。

这个数量被看作进化的距离（其准确值依赖于分子进化模型的选择），以这些距离构造矩阵，即距离矩阵。

然后，运行一个聚类算法，即从最相似（也就是之间距离最短）的序列开始，通过距离矩阵计算出实际的进化树；或者通过将总的树枝长度最小化而优化出进化树。

**简约法 （parsimony method)**

最大简约法：要求用**最少的改变**来解释所研究的**分类群之间的差异**。

对一个数据集（序列组）可能产生多个最简约树， 这时就必须将它们综合起来，产生一致树。

进化简约法（加权最大简约法）：

考虑实际替换的代价不同，如转换和颠换的区分**。**

**最大似然法 （maximum likelihood method）**

**评估所选定的进化树能够产生实际观察到的数据的可能性：（基于置换）**

**针对一个位点的进化，先把某种组合的核苷酸置于进化树的内部结点，根据取代函数计算每一段进化的可能性，将所有段的这种可能性相乘，得到此进化树以此组合为进化途径产生此位点数据的可能性（似然值）；**

**换组合，再算。将所有组合的这种可能性相加，得到此进化树产生此位点数据的似然值；**

**换位点，再算。将所有位点的这种可能性相乘，得到此进化树产生整个序列组数据的似然值。**

**换进化树，再评估，相比较。具有最大似然值（产生此数据可能性最大）的进化树被认为是最可能的实际进化树。**

**（加，乘关系）**

例如，转换出现的概率大约是颠换的三倍。在一个三条序列的比对中，如果发现其中有一列为一个C，一个 T和一个 G，我们有理由认为，C和 T所在的序列之间的关系很有可能更接近。由于被研究序列的共同祖先序列是未知的，概率的计算变得复杂;又由于可能在一个位点或多个位点发生多次替换，并且不是所有的位点都是相互独立，[概率计算](https://baike.so.com/doc/733047-776056.html" \t "https://baike.so.com/doc/_blank)的[复杂度](https://baike.so.com/doc/439166-465040.html)进一步加大。尽管如此，还是能用客观标准来计算每个位点的概率， 计算表示序列关系的每棵可能的树的概率。 然后，根据定义，概率总和最大的那棵树最有可能是反映真实情况的系统发生树

结果：（一般来说）

A.进化速率恒定时：简约法≤距离矩阵法；

最大似然法依赖于进化模型。

B.进化速率可变时：简约法≤距离矩阵法＜最大似然法。

计算时间：距离矩阵法<简约法<最大似然法

序列间有极高相似性：简约法。

序列间有较明显的相似性：距离矩阵法。

序列间没有较明显的相似性：最大似然法。

自展法：Bootstrap又称自展法，是用小样本估计总体值的一种非参数方法，在进化和生态学研究中应用十分广泛。例如进化树分化节点的自展支持率等。  
  
Bootstrap的思想，是生成一系列bootstrap伪样本，每个样本是初始数据有放回抽样。通过对伪样本的计算，获得统计量的分布。例如，要进行1000次bootstrap，求平均值的置信区间，可以对每个伪样本计算平均值。这样就获得了1000个平均值。对着1000个平均值的分位数进行计算， 即可获得置信区间。已经证明，在初始样本足够大的情况下，bootstrap抽样能够无偏得接近总体的分布。

自展法利用对原始数据随机抽样产生的自展数据获得大量“虚拟”树，然后检验这些“虚拟”树对被评估树的支持程度。

序列组——构树（一个树）—— 自展法检验

自展法已由一种检验方法发展为构树的基本步骤。

序列组——自展——构树（大量虚拟树）——一致树

国际上常用的系统树构建软件包为PHYLIP和PAUP，PAML、MEGA、其中PHYLIP发布最广、用户最多

CHAPTER6

基因组特性，根源性和完备性，遗传多态性

结构基因组分析 (基因组作图，STS \*)

功能基因组分析

比较基因组分析 (COG \*)

基因组信息资源 (NCBI，UCSC\*，Ensembl)

基因组是指细胞生物体的单倍细胞、细胞器或病毒所包含的所有基因与基因间序列的总体。意义在于根源性（生命形式的“实质载体”）完备性（生命体的全部信息）

基因组学的研究有三方面

1. 基因组测序：基因组有什么样的序列？

2. 基因组解析：基因组各部分是什么？它们具有什么样的生物学意义？

3. 基因组比较：基因组是怎样进化的？

生命和信息的内在联系：信息来源于选择，生物信息来源于自然选择。

• 机会：基因组序列信息的丰富性为生物信息 学提供了很好的研究对象；

• 挑战：基因组所含信息量至少比单个基因高 几个数量级；

• 意义：基因组相关的分析赋予了生物信息学更高的意义根源性和完备性）。

基因组大小与进化位置并不绝对相关人基因组3.3G无恒变形虫670G （C值悖论）

非编码区的意义还不很明确，现在已经成为研究热点之一。

—— 目前的转录组研究表明，有很多非编码区序列实际上转录为RNA，功能尚不十分清楚。

—— 完全“无用”的序列为沉默突变的积累提供了一个平台。这应该是真核生物能够进化到如此复杂的一个重要原因

**基因组中的重复序列**

**• 高度重复序列**

**重复几百到几百万次，一般少于10个核苷酸残基。如着丝粒、 端粒序列。~10-20%。**

**• 中度重复序列**

**重复十次到几百次。如rRNA基因、tRNA基因、转座子序列和 某些蛋白质（如组蛋白、肌动蛋白、角蛋白等）的基因。~20%。**

**• 轻度重复序列**

**重复二到十次。如一些组蛋白基因。~10%**

**• 单拷贝序列 ~50%-60%。**

高度重复序列：

小卫星（minisatellite）,散在。

微卫星（microsatellite）,着丝粒。

端粒序列：染色体末端（和年龄有关）。

中度重复序列：

编码rRNA的基因，核仁组织区。

转座子序列(SINE，LINE）,散在。

单核苷酸多态性（SNP）：是指基因组内特定核苷酸位点上存在不同的碱基。

（考虑这样一个事情，收集全世界人类的基因组信息，发现人类之间的基因组相似度很高，有99.9%以上，也就是说绝大部分位点的核苷酸都是相同的，只有剩下的一小撮位点不同，请注意，剩下的一小撮其实是分散在整个基因组里，也就是这个位点可能不同，隔一大段后又发现一个位点不同。这些不同的位点大部分由SNP组成，也就是说是单个位点的不同，而非连续一串的不同。另外，SNP的指标是这种不同要在整个群体中达到1%以上才可以。例如某条染色体上的770号位的核苷酸有98个人是A，剩下1个人是T，1个人是C，这就是一个SNP。SNP可能会造成沉默突变或者有义突变，但是大部分SNP位于非编码区。某个SNP突变可能会造成对疾病的易感性，因此可以利用SNP做疾病分析。）

简单序列重复多态性（SSRP）：卫星DNA (微－STR，小－VNTR） ，重复次数可变

**结构基因组分析，基因组作图**

基因组水平的研究最重要的特征之一是有位置信息 ——

• 基因组序列图：什么位置是什么核苷酸？

• 基因组元素图：什么位置是什么元素 （elements：基因、疾病倾向位点、标记等）？

——终极目标：把所有图依照位置对应起来。

基因组元素图有Cytogenetic maps

Genetic linkage maps

Radiation Hybrid maps

STS maps

Clone-based maps

**基因组测序策略**

Whole-genome shotgun (WGS) :

一次性全部打断（机械法或限制性内切酶法）；对每个片段两端测序；用计算机程序聚类这些片段，拼接为contigs；进一步拼接为genome。适用于小型基因组；或配合HS用于大型基因组。

Hierarchical shotgun (HS) :

首先打断为contigs（BAC克隆等，已知这些contigs拼接为genome的方式）；再对一个contig打断，对各片段两端测序，再用程序拼接得到这些contigs的序列信息。适用于大型基因组。

Contigs

定义:

Contigs（连续序列）是通过测序多个重叠的DNA片段并将它们正确地拼接起来得到的一长串没有间断的序列。在基因组测序中，一个contig代表基因组中连续的、正确顺序的DNA序列。

在HS中的作用:

在分层枪式测序中，整个基因组首先被分解成更小的片段，这些较小的片段测序后形成了contigs。这些contigs是基因组重组的基础单元，理解和确定它们的顺序及位置是整个基因组拼接过程的关键。

BAC克隆

定义:

BAC（Bacterial Artificial Chromosome）克隆是一种大分子量DNA克隆方法，用于在细菌中复制大片段的DNA。BAC克隆可以容纳约100-300千碱基对的DNA序列，这使它成为存储和操作大片段基因组DNA的理想载体。

在HS中的作用:

在分层枪式测序策略中，BAC克隆作为大型基因组的一个管理单位。首先，研究人员会将整个基因组分解成许多较大的片段，并将这些片段插入到BAC载体中，每个BAC克隆代表基因组中的一个较大区域。然后，这些BAC克隆被进一步打断成更小的片段用于测序。

分层枪式测序的流程

分解基因组:

首先，研究人员将整个基因组打断成较大的片段，并插入BAC载体中，形成BAC克隆库。这些BAC克隆各自包含基因组的一个小区域。

选择和测序BAC克隆:

然后，从BAC克隆库中选择特定的BAC克隆进行工作。每个选定的BAC克隆会被进一步打断成更小的片段，这些片段通常会在两端进行测序（称为末端测序）。

组装Contigs:

通过末端测序产生的读序信息，以及已知的BAC克隆信息，研究人员可以将这些小片段重新组装回较大的contigs。这些contigs反映了原始BAC克隆中DNA的序列。

拼接整个基因组:

最后，通过确定不同contigs之间的正确顺序和位置，可以将它们拼接成完整的基因组序列。

**EST序列数据库**

Genbank的EST部分 （dbEST）收录了向公众公开的部分EST数据（专门提交方式，注释少） ；

TIGR （The Institute for Genomic Research）免费提供公众使用。其中的HGI（Human Gene Index）令人关注。

WashU-Merck 来源于人类各种组织的规范化cDNA文库（见书表4－4），免费提供公众使用

**序列的拼接**

对于不重叠的片段找延伸信息，重叠片段寻找一致性序列信息。

拼接软件：

phrap、Assembler、sequencher等。

（EST： phrap、Cap3、Cat、Staden、TIGR等）。

（注意，现在的生物学实践中已经不再使用EST方法）

**STS map**

**An STS is a relatively short sequence (200 to 500 bp) occurring only once in the genome.**

路标是种内基因组间保守性超过基因组内保守性的短序列，若根据全相似性，则路标太少。而**根据两端特异相似加上中间长度的大致限制，可以找到很多路标，STS就是这种路标。**于是找STS路标必须根据这个特性，即PCR或e-PCR，而不是根据全相似性的杂交或电子杂交（BLAST)。

（STS的用意是，拥有了STS与基因组的映射关系之后，例如，你知道某个STS1在第一条染色体的100位置。你可以通过STS寻找你想要的基因所在的位置，例如基因A可能在STS3和STS4之间，而STS3和STS4分别在第二条染色体的1000和2000位，因此A在这中间）

STS来源：序列已知的genes（/EST）、random genomic DNA —— Bioinformatics, A practical guide to the analysis of genes and proteins, 2nd version. P27. 另外，还可来源于SSRP。其中cDNA的3‘-UTR，及SSRP很可能在基因组中唯一，原因是它们不保守，变异快。

确认STS：用PCR（在所有基因组序列都存在的情况下，唯一P出）；在基因组序列已知后，可用reverse-ePCR。

定位STS：在基因组计划完成前，RH (radiation hybrid) 法对STS的定位有作用，但完成后，用reverse－ePCR的方法就可以对STS进行精确定位。

STS, actually, represents the short sequences that can be specifically amplified by PCR in the presence of all genomic sequences --- operationally, an STS is defined by a pair of oligonucleotide.

In some cases, the size (/ sequence) of the amplified PCR product may be polymorphic

We could identify STSs in the target DNA sequence by searching for subsequences that match the PCR primers and are in the correct order, orientation, and spacing of a real PCR

（简要翻译：从操作上讲，一个STS并非由STS序列定义的，而是由STS序列两边的引物（寡核苷酸）定义的，因此在不同个体之间STS不一定相同）

模拟PCR —— E-PCR

BLAST is more analogous to “electronic hybridization”. It does not focus on the PCR primers alone. E-PCR focuses on the PCR primers alone.

（普通BLAST是一种模拟电子杂交，并不要求引物的一致性，而我们想要的是引物不变，从而得出STS，因此需要模拟PCR）

**基因组解析：从传统模式到后基因组时代（基因组功能）**

**功能－蛋白－基因－基因组（结构基因组分析）**

**基因组－基因－蛋白－功能（功能基因组分析）**

基因组注释主要是基因预测，在DNA序列分析中已经讲过，这里放在基因组范围来理解。另外，还涉及很多基因间序列的内容。（即对基因组进行注释，哪里到哪里是启动子，哪里是CDS序列等等这样的注释，更多的注释包括基因间是否有相互作用等）

Genome-wide association study ---- 可能直接从基因组到功能。(snp…第二代测序）。

--- the genome-wide study of the function of DNA, as well as its RNA and protein products, to understand the relationship between genotype and phenotype.

--- by analysis of

SNP and epigenomics, (DNA  function)

expression variation, (RNA, proteins  function)

molecular interactions, (RNA, proteins  function)

natural and experimental disruptions (disfunction  function)

（总的来说功能基因组分析是分析基因型和表型之间的关系，分析基因对应的功能）

从广义上来看，功能基因组是包括转录组和蛋白质组的。

**比较基因组分析**

**（由于可用的基因组多起来了，自然推动了比较）**

**Cluster of orthologous groups（COG）——直系同源体簇 （基因组比较分析）**

**The availability of multiple complete genome sequences spurred the demand for the construction of an evolutionary classification of genes from these genomes，which appears to be a natural framework for comparative genomics.**

**直系同源（orthology）：由“物种分化”而产生， 一般有功能一致性；可以反映物种血统上的同源性，即物种进化的历史。**

**比如，人的肌红蛋白和小鼠的肌红蛋白。**

**并系同源（paralogy）：由基因“多重化（duplicating）＋功能分化” 而产生。**

**比如，人肌红蛋白和人的血红蛋白中的α、β亚基**

**（一个簇是将多个物种中的基因根据它们的直系同源关系分组。这些簇中的基因通常执行相似的功能，并且来自共同的祖先基因。可以类比为一个等价类）**

这是概念上的理解，即通过功能来界定，问题1.对于远缘序列，功能可能都不同；问题2.实际上有时候，我们对研究基因的功能并不十分了解，或者说有的了解，有的不是很清楚，在genome信息数据先于功能研究的后基因组时代尤其如此。对于问题2，一种传统的做法是通过相似性来判断，即直系同源更相似（因为功能相似而被自然选择保护，而并系同源本身就是“…功能分化”而产生，更倾向不同）；但远缘基因组中的直系同源的相似性低（功能可能已经不太相同了）于并系同源，也就是这种方法也无法解决问题1。 所以有了COG的概念——干脆把不同基因组中的所有同源基因都归在一起。Each COG is assumed to have evolved from an individual ancestral gene through a series of speciation and duplication events.

Orthologs are homologous genes derived by vertical descent from a single ancestral gene in the last common ancestor of the compared species. Paralogs, in contrast, are homologous genes, which, at some stage of evolution of the respective gene family, have evolved by duplication of an ancestral gene.

Orthologs typically occupy the same functional niche in different species, whereas paralogs tend to evolve toward functional diversification.

识别直系同源不仅对于建立物种树十分重要，还对新基因功能预测十分重要。（一般我们是通过假设直系同源序列最相似来判断。经典看法）

小鼠肌红蛋白和人血红蛋白的α或β亚基是什么同源呢？当并系同源的相似性更高，我们可将其归为一簇，和别的物种中的相应簇称为直系同源。

如果肌红蛋百和血红蛋白亚基之间非常相似，比如其相似性高于不同物种间肌红蛋百的相似性，这种直系同源体簇的划分就十分重要，而小鼠肌红蛋白和人血红蛋白的α或β亚基就可看作直系同源，因为它更反映的是物种分化，而不是基因多重化带来的功能分化。

（以上标绿部分表示在新版ppt以及课堂上已经没有讲过）

**全基因组进化分析还可能用到基因组特性，如G+C含量，二核苷酸、三核苷酸（密码子）、四核苷酸等的频率等可能和物种特异相关的信息。**

**宏基因组分析**

实际上，宏基因组（或微生物组）构成了一种“环境因子”，和温度、湿度及pH等类似，而我们现在利用DNA测序来标征和解析这种因子。~对人微生物组和自然环境微生物组都适用。

**转录组分析：**

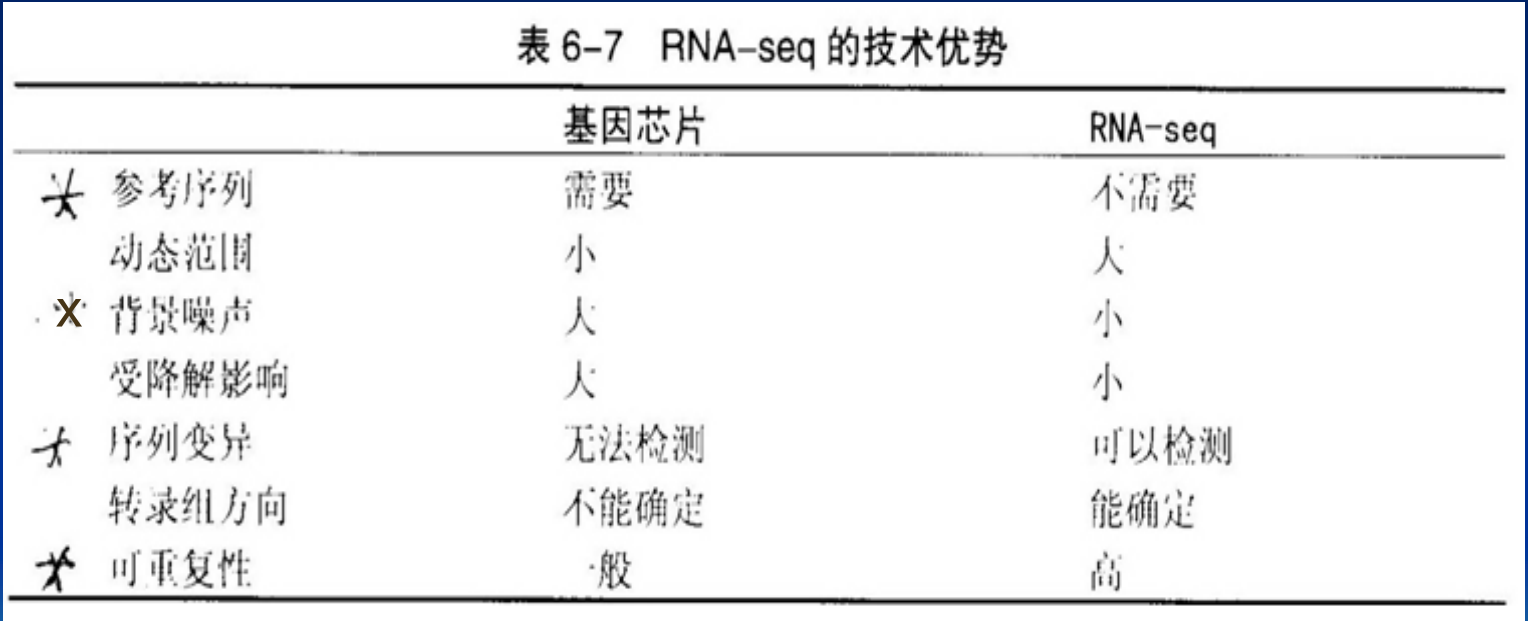
**Microarrays ——**本质是生物信息的**集成性平行分析**：利用**核酸分子杂交**（蛋白质分子亲和）原理，通过**荧光标记可视化**，借助计算机分析处理，可迅速获取大量生物信息，效率是传统手段的成百上千倍**。也称** biochips，microchips, DNA microarrays, DNA arrays, DNA chips, gene chips, others. （在一个芯片上划分很多小格子，每个小格子放入不同的序列的大量荧光标记的克隆。加入欲识别的序列后，mRNA落入每个小格子，如果匹配就被杂交显色，如果不互补配对则显示另一种颜色）

Two major technologies ---- Spotted DNA microarrays Oligonucleotide GeneChips

Microarray分析：图像分析（去噪音和信号数据化）、标准化（重复实验的可比性）、Ratio分析（两色荧光的比值）、基因聚类分析（寻找同类基因）

（事实上，DNA array技术也已经较老，在2023年的当下已经被RNA-seq所取代了）

**RNA-seq** describes a collection of experimental and computational methods to determine the identity and abundance of RNA sequence in biological samples. （RNA-seq能够识别转录组中有哪些RNA，已经各个RNA的转录水平如何）



**(考了上图)**

RNA-seq methods are derived from generational changes in sequencing technology.

RNA-seq 分析的靶标并不局限于mRNA， 还包括noncoding RNA，如tRNA、 rRNA 、miRNA和lncRNA等，是真正的、全面的 转录组分析技术。

RNA-seq中的cDNA文库只是cDNA双链片段的集合，没有克隆 ---- 只符合抽象意义上的cDNA文库的概念。

由于概念和实验方法都相近，RNA-seq的结果可以看作升级版的EST，尤其是辅助基因发现时（向核苷酸数据库提交时）。

之所以EST不放在转录组分析中讲，而RNA-seq放在这里讲，因为表达差异分析才是转录组分析的主要任务，而辅助基因发现只是辅助基因组分析的方面。

相关分析：质量控制、读段清理、序列拼接、表达定量、**相关系数分析、表达差异分析**

**EST和RNA-seq都是取样并测序转录组（用cDNA），其主要不同在于：**

EST RNA-seq

文库构建方法 克隆 不需要克隆（PCR就行了）

// 个例结果 可回溯 无法回溯

测序方法 第一代 第二代

读段 少 多多

表达差异分析 粗放 精细

主要目标 定性（辅助基因发现） 定量（表达差异分析）

在表达基因中 侧重基因 侧重表达

（它们都还可以用于可变剪切研究等）

（注意：我在看这一部分时有些疑惑什么是克隆，构建文库为什么要用克隆，为什么不用大家熟知的PCR。首先，克隆，回忆高中学的单克隆抗体技术，是将基因插入载体后导入宿主细胞（细菌），什么限制酶啊，之类的东西，利用宿主细胞的自我复制达成克隆的目的从而构建cDNA文库。那是因为EST已是上世纪的技术，没有好的PCR方法，然而在当下，直接PCR即可）

**DNA分子克隆技术**（也称基因克隆技术） ：在体外将DNA分子片段与载体DNA片段连接，转入细胞获得大量拷贝的过程中DNA分子克隆(或基因克隆)。其基本步骤包括：制备目的基因→将目的基因与载体用限制性内切酶切割和连接，制成DNA重组→导入宿主细胞→筛选、鉴定→扩增和表达。载体（vecors）在细胞内自我复制，并带动重组的分子片段共同增殖，从而产生大量的DNA分子片段。主要目的是获得某一基因或DNA片段的大量拷贝，有了这些与亲本分子完全相同的分子克隆，就可以深入分析基因的结构与功能，随着引入的DNA片段不同，有两种DNA库，一种是基因组文库（genomic library）,另一种是cDNA库。

PCR是聚合酶链反应的英文缩写。该反应是根据DNA变性，复性原理设计的。反应系统中包括含目的基因的微量样品，耐热的DNA聚合酶，4种dNTP（脱氧核苷酸）以及两种过量的引物。引物一般长20-30bp（可以用人工合成）。该反应分三步进行： 　　a.DNA模板（目的基因和引物）高温度变性（95°C） 使之解链。 　　b.降温复性（40°C-60°C），使目的基因与引物杂交。 　　c.在70°C-75°C的温度下，DNA聚合酶进行酶促聚合反应，使目的基因DNA序列在3’位置上逐渐延伸。新合成的引物延伸链又可以作为下一步反应的模板，如此反复循环进行，按照2^n的指数扩增，短时间内可获得大量的DNA目的基因。

（标绿部分未讲过）

**第二代测序 （SGS/NGS）: 通量高**

Illumina: 600G nt / run (Sanger: 0.1M nt / run) ;

Roche 454: 700nt / read (似Sanger法).

SOLID: sequencing by oligonucleotide

ligation – accuracy 99.99% (Sanger：99%) ;

**第三代测序（TGS）: 单分子测序**

Pacific Biosciences: 5000 nt / read ;

Nanopore technologies: 微电流测序 – …RNA/P 微电流测序：离子经过一个膜上的蛋白孔（纳米级），产生微电流，当DNA同时穿过此孔时影响此电流，不同的核苷酸的影响不同

**RNA-seq有取代基因芯片的趋势**

**从转录基因组到转录组：RNA-seq 分析的靶标并不局限于mRNA，还包括noncoding RNA，如tRNA、 rRNA、miRNA和lncRNA等，是真正的、全面的转录组分析技术。**

CHAPTER7

蛋白质组和蛋白质组分析

蛋白质组表达分析

分离——双向凝胶电泳

鉴定——质谱分析（肽质指纹分析\*)

蛋白质组网络分析

蛋白质组信息资源

蛋白质组定义：在某种内在和外在条件下，一个基因组表达产生的所有蛋白质的总体。（具体。

（细胞中所有蛋白质的集合被称为其蛋白质组。不同于相对不变的基因组，动态的蛋白质组会随着成千上万种内外细胞环境信号的变化而在几分钟内发生改变。）

蛋白质组特性：

整体性：不同于传统的蛋白质化学，其研究对象不是单一或少数的蛋白质，更强调整体性，从而有利于对生命活动机制的全面了解。

多态性：对于多细胞生物，组织特性不同，蛋白质组不同。

动态性：蛋白质组随外在和内在条件而变化，如E.coli随环境不同；发育，病理情况。

蛋白质组分析是生物信息学的又一重要应用方向：信息的丰富性和复杂性——机会与挑战。

相关研究也赋予生物信息学更高的意义（直接面向细胞全面功能，以系统生物学为目标）。 **分离——双向凝胶电泳**

图像分析：斑点位置和密度分析（包括斑点分离、背景消减等工作）。

斑点配比：不同图像中同种斑点的识别（蛋白差异表达分析的前提） 。

聚类分析：蛋白质表达矩阵。

（第一向等点聚焦，第二向SDS-PAGE）

**鉴定——质谱分析（肽质指纹分析\*) （传统方法如Edman法已经不适用）（可能考相关的翻译题）**

**质谱分析的工作原理：**

**样品准备：首先将样品准备成适合电离的形式。对于不同类型的样品，可能需要不同的准备方法。**

**电离：样品分子被电离成带电粒子。电离方法有很多种，包括电喷雾电离（ESI）、基质辅助激光解吸/电离（MALDI）等，选择哪种电离方法取决于样品的性质。**

**质量分析：带电粒子进入质谱仪的质量分析器，根据它们的质荷比被分离。常见的质量分析器有四极杆、飞行时间分析器（TOF）和离子阱等。**

**检测：分离后的离子被检测，通常通过测量它们撞击检测器产生的信号强度。**

**数据分析：最后，所得到的质谱图（即离子强度vs.质荷比的图）被分析，用以识别样品的组成或确定某些物质的质量。**

**（请回忆高中电磁学相关内容，洛伦兹力不做功）**

**肽质指纹则指将一段多肽按特定酶切位点进行酶解后放入质谱仪中获得质谱图，质谱图就是一份指纹。**

**可能出现肽质指纹相同的不同蛋白质吗？**

**理论上可以，实际上不太可能——所以叫指纹。因为蛋白质的序列**

**特异性要求太严格了，两个完全不同的都有功能的，但肽质指纹特征相同的序列，实在不太可能。**

**具体方法：（请注意big picture，在这一步中，我们分离了某个未知的蛋白质，想要知道它的序列）**

酶解拟鉴定蛋白质，测定实际肽段质谱图；

从蛋白质数据库中随机挑选一段蛋白质序列，理论酶解之，并计算出各个肽段的分子量，从而构建出理论质谱图；

将实际肽段质谱图与此蛋白序列的理论质谱图进行比较，进行相似性计分；

从数据库中重新挑选一段序列，重复以上两步；

最后，根据相似性计分确定最可能的蛋白序列。

该方法的延展还可以分析是否有磷酸化等修饰，因化学修饰后荷质比将改变。

传统方法：

Edman法——工作量大（无法高通量）；

aa组分分析—— 比较“粗放”。

现代方法（质谱技术的应用）：

肽质指纹分析（Peptide-mass fingerprinting）；

串联质谱（tandom Ms or Ms/Ms) 。

蛋白质组相关数据库：双向凝胶电泳图： **Swiss-2DPAGE**，World -2DPAGE List……

蛋白质相互作用：BIND (Biomolecular Interaction Network), **DIP (Database of Interacting** Proteins), GRID (General Repository for Interaction Datasets)……

生化代谢网络分析-系统生物学

研究两方面内容

1. 两个蛋白质是否相互有接触作用？（interaction）
2. 两个蛋白是否相互有化学反应？（reaction）

生化路径数据库： KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway, UniPathway, EcoCyc

**ExPASy**

**综合性蛋白质信息服务器，**苏黎世生物信息学研究所建立；

**最完备的蛋白质数据库网站之一，包括Uniprot (主要是SWISS-PROT/TrEMBL)、PROSITE（蛋白质家族和编码区数据库）等公共蛋白质数据库，面向蛋白质组，并已逐步转向综合性的生物信息学服务器。**

**ImageMaster肽质指纹分析软件**

CHAPTER8

序列通过折叠为结构而最终形成功能是使生命得以实现的重要自然原理。

domain不仅是一个序列相关的概念，也是一个结构相关的概念。

一个domain可以是一段保守的序列，可以是一个独特的三维结构，可以是一个序列或结构拥有特定的功能。

催化功能——形状＋特定位置的特定aa

结构的保守性可能超出序列比较的可识别性

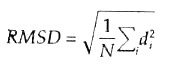
序列比对：源序列与目标序列之间按残基位置相对排列。使序列之间的相似程度最大。

结构比对：源结构和目标结构之间按残基位置相对排列，使结构之间相似性程度最大——相应残基的α碳原子空间位置最接近。

（结构相似不一定序列相似）（同源进化一定序列相似，而结构相似很可能是趋同进化）

Root Mean Square Deviation (RMSD，均方根误差（这种定义在物理化学课程中也有涉及），相应残基α碳原子平均空间距离的度量 .结构相似性衡量的重要指标

RMSD<1.5 埃 —— 非常相似



VAST（vector alignment search tool）NCBI里结构比对查询的工具

蛋白质结构源数据库---PDB （勿与序列复合数据库UniProt搞混）

由于结构的高度保守性，基于结构的蛋白质分类方法可以更准确地反映蛋白质的功能和演化关系。即使在序列相似性不高的情况下，结构相似性仍然可以揭示蛋白质之间深层的演化联系或功能类似性。

Families —— 近缘同源（序列相似性较明显）；

Super-families —— 远缘同源；（有些结构相似蛋白质的序列几乎看不出序列相似性——远缘同源）。

Super-folds —— 有一定相似但不一定同源.

**蛋白质结构分类数据库**

**SCOP** (Structure Classification Of Proteins)

依结构相似性进行分类，等级细化顺序为——Class，Folds、 Gene superfamily 、Gene family。

**CATH** (classification by Class, Architecture, Topology, and Homology)

依结构相似性进行分类，等级细化顺序为——Class、Architecture、Topology、Homology superfamiliy、Sequence family。

**FSSP** (Fold classification based on Structure-Structure alignment of Proteins) or (Families of Structurally Similar Proteins)

SARF (Spatial ARrangement of backbone Fragments)

CASP (Critical Assessment of Structure Prediction) competitions一个有关蛋白质结构预测的竞赛，人工智能的表现突出（alpha-fold）

蛋白质结构预测相对于RNA结构预测更复杂：因为aa之间的相互作用远比碱基配对要复杂，另外关于蛋白质的预测的主要是三级结构，而RNA主要预测二级结构。

**方法：**

1. Ab initio（从头预测）: Ab initio prediction 原理自由能最小

Ab initio prediction: 基本物理学原理－－量子力学、统计热力学，计算最小自由能。系统太大，比较困难。但逐渐在预测竞赛中崭露头角。这种方法本身令人感兴趣，因为它完全依赖于序列信息，不需要其它结构信息，是对科学界的挑战。如果能很好地实现，将是巨大的科学成就。（是数学上的珍宝，却不是当今最实用的方法）

（2）Knowledge-based: Comparative modeling （Homology modeling同源模建）\*Theoretical basis：Sequences with more than 25% identity over an alignment of 80 residues or more adopt the same basic structure.在80个或更多个残基的比对上具有超过25％同一性的序列采用相同的基本结构。寻找序列相似物==>结构。70%以上，几乎和实测结构完全一致；70%以下，构建结构的质量逐渐降低（有说法：每10%相似性的drop，带来0.3埃的误差），这时候往往需要通过手工介入（如关键aa必须对应上），特别是40%以下，如果无手工介入，可能“死得很惨”。

一般来说，30％以下的的相似性，不能用这种方法。25%（同源模建的底限）

比较建模的基本步骤通常包括：

模板识别：通过将目标蛋白质的序列与已知结构的蛋白质数据库进行比对，来识别一个或多个合适的模板。这些模板是已经解决了结构且与目标蛋白质有较高序列相似性的蛋白质。

序列比对：将目标蛋白质的序列与选定模板的序列进行对齐，确定它们之间的相似和差异区域。这一步骤对于后续准确建模至关重要。

模型构建：基于序列比对，将目标蛋白质的序列结构“映射”到模板结构上。这通常涉及到构建蛋白质的主链，然后添加侧链，最后调整以获得最佳构象。

模型优化和评估：构建的模型需要通过能量最小化等方法进行优化，以确保其物理合理性。然后，通过各种评估方法检查模型的质量，包括其立体化学质量、表面特性、模板与模型之间的结构一致性等。

模型验证：通过比较模型预测结果与已知的实验数据或使用特定的验证工具来评估模型的准确性和可靠性。

Fold recognition（Threading）和同源模建类似，也限于已知的蛋白质结构，但原则上并不要求序列十分相似，如可低于25%（同源模建的底限）。

蛋白质结构预测的“线性穿梭”（Threading）技术，是另一种用于预测蛋白质三维结构的方法，特别是当目标蛋白质的序列与已知结构的蛋白质序列相似性不高时使用。它基于这样一个假设：即使蛋白质序列之间的相似性很低，它们可能仍然共享相同的蛋白质折叠模式（fold）。因此，即使传统的同源模建方法不适用，我们仍然可以尝试预测这些蛋白质的结构。

借助更远缘序列的相似性来推测结构

（3）Knowledge-trained: Secondary structure prediction Predicting the conformational state of each residue in three categories, helical, strand, and coil，usually based on ideas reflecting the preference of a residue for a particular secondary structure.

Accuracy: (early) 60%, (+conserved domains) 66%, (+structural data + sophisticated algorithms) >70%.

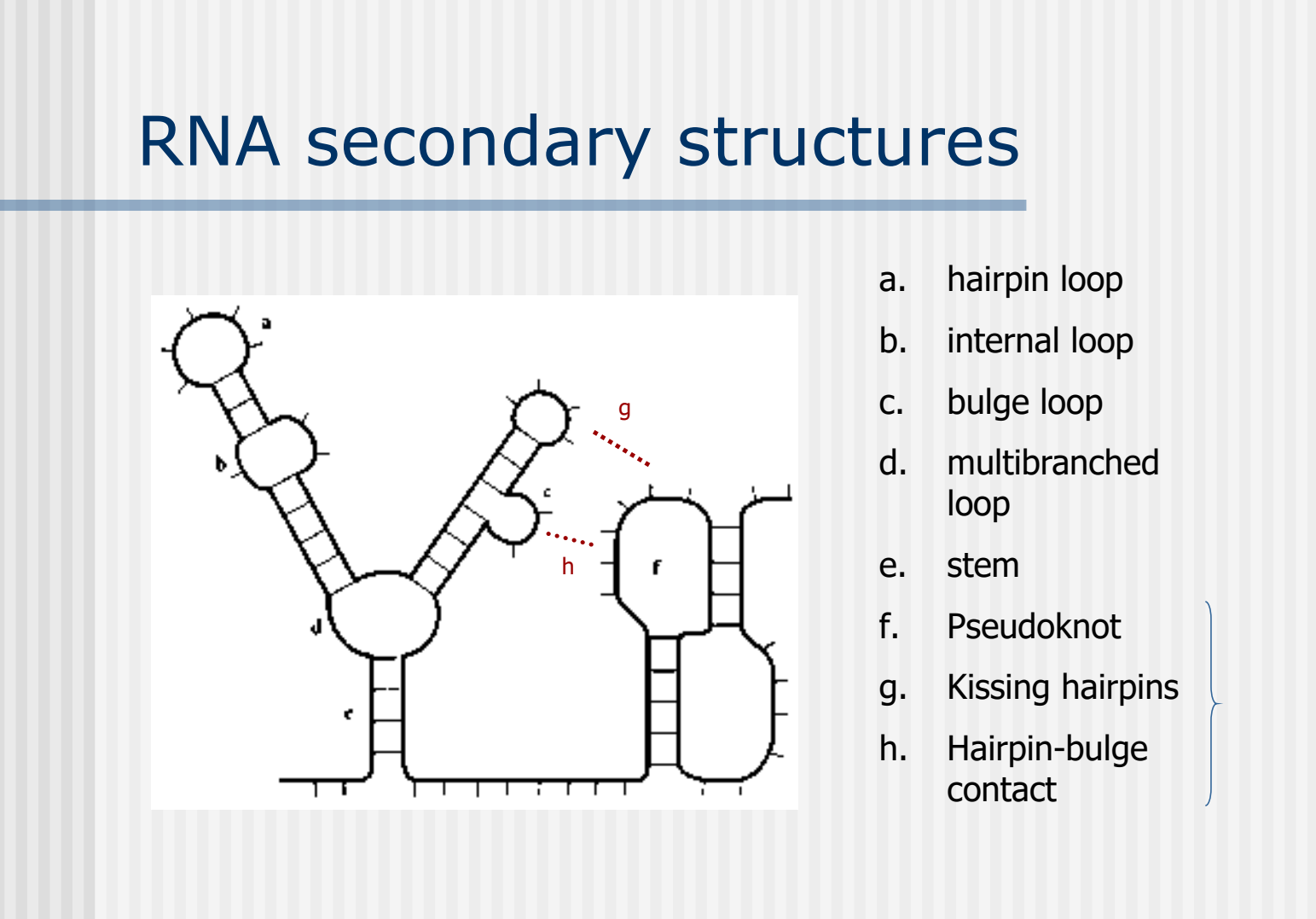
Not for integral membrane proteins（需要专门的算法）.

人工智能方法也属于此类。

SWISS-MODEL是一个ExPASy数据库下的一个工具即软件，寻找蛋白结构的同源模型。

（在2023年的如今，已经没人用上述传统方法预测了，全部使用人工智能方法预测蛋白结构）

RNA结构分析主要是结构预测，没有结构比对及结构分类的部分。mRNA: 转录和翻译调节中的作用：



1. 从头预测
2. "Evidence-associated prediction"（证据相关预测）通常指的是结合实验数据或其他类型的证据来预测RNA结构的方法。Covariation prediction（共变异预测）是一种RNA结构预测方法，尤其用于预测RNA分子的二级结构。共变异预测基于这样一个观察：在RNA分子中，如果两个核苷酸在演化过程中共同发生变异（即它们的序列同时改变），那么这两个核苷酸很可能是互相配对的，因为配对的核苷酸在结构上是相互依赖的，任何变化都需要保持配对的互补性以维持整体的结构稳定。
3. 遗传算法
4. 特定序列同源建模
5. 动态规划算法

RNA结构预测方法，相应软件Mfold；Sfold; Vienna RNA;RNAstructure;

Ab initio prediction： Choosing the most energetically stable sets from the many possible choices of complementary sequences that can potentially base-pair.

Limitation: only local effects; Pseudoknots --- computation-costly.

Efficiency and accuracy: 70% structure for a sequence.

Covariation prediction

Genetic algorithm

Specific sequences with known structures ~ databases

动态规划算法：Dynamic programming: 从局部最优延伸到全局最优——假设局部不受全局影响；如果局部受全局影响则不一定能得到全局最优。对于全局序列比对，这种方法真的可以找到最优解（主要原因是序列不能颠倒）

遗传算法是一种可行算法。找到较好的可行解而不保证最优解。（就像人类进化一样）

